(12) 公開特許公報 (A)　(11)特許出願公開番号　特開平11－272699
(43)公開日　平成11年(1999)10月8日

BD

識別記号	FI	
./30		
17/27	G06F 15/401	320A
	15/20	550A
	15/38	D
	15/40	370A

審査請求　未請求　請求項の数17　OL　(全 74 頁)

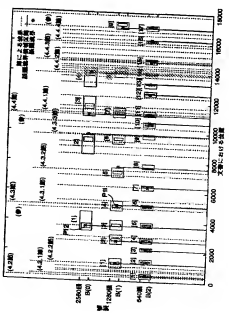
(21)出願番号　特願平10－72724
(22)出願日　平成10年(1998) 3月20日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　仲尾　由雄
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　弁理士　大菅　義之　(外１名)

(54)【発明の名称】　文書要約装置およびその方法

(57)【要約】
【課題】　文章一般に見られる現象をもとに、文書中の話題構成を自動的に認定して、話題構成に対応する要約を作成することが課題である。
【解決手段】　文書要約装置は、異なる大きさの複数の窓を用いて、文書中の各位置における語彙的結束度を計算し、各話題の階層毎に話題境界の候補区間を求める。次に、異なる階層の候補区間を順に統合していくことで、各階層毎に話題境界を認定する。そして、要約作成対象の話題のまとまりと、それを含む大きな話題のまとまりとの関係に基づき、重要文を抽出して要約を作成する。

話題構成の第1の認定結果を示す図



1

【特許請求の範囲】
【請求項１】　与えられた文書中の話題の階層的構成を認定する構成認定手段と、
認定された各話題に関する重要語を抽出する抽出手段と、
前記重要語の出現状況に応じて、各話題のまとまりから重要文を選択し、該重要文を用いて要約を生成する選択手段と、
前記要約を出力する出力手段とを備えることを特徴とする文書要約装置。
【請求項２】　前記構成認定手段は、前記文書中の各位置の近傍領域における語彙的結束度を求め、該結束度に基づいて話題境界を認定し、該近傍領域の大きさを段階的に縮小しながら話題境界の認定を繰り返すことで、大きな話題のまとまりから小さな話題のまとまりに至る話題の階層的構成を認定することを特徴とする請求項１記載の文書要約装置。
【請求項３】　前記構成認定手段は、前記各位置の近傍領域における前記結束度を、各位置の前後に設定した２つの窓の中に含まれる語彙の類似性から求め、窓幅を段階的に縮小しながら前記話題境界の認定を繰り返すことを特徴とする請求項２記載の文書要約装置。
【請求項４】　前記構成認定手段は、前記結束度を移動平均した値を、移動平均の開始点における右結束力および移動平均の終了点における左結束力として扱い、右結束力と左結束力が拮抗している位置の近傍を話題境界の候補区間と認定する候補区間認定手段を含み、該候補区間を用いて話題境界を認定することを特徴とする請求項２記載の文書要約装置。
【請求項５】　前記結束度の比較的小さな領域を処理対象から除外して、該結束度の比較的大きな領域を重要部分として抽出する重要箇所特定手段をさらに備え、前記選択手段は、該重要部分に対応する話題のまとまりから前記重要文を選択することを特徴とする請求項２記載の文書要約装置。
【請求項６】　前記抽出手段は、前記話題のまとまりの範囲に出現する語彙が、該話題のまとまりに対して特徴的であるかどうかを評価し、評価結果に基づいて前記重要語を抽出することを特徴とする請求項１記載の文書要約装置。
【請求項７】　前記抽出手段は、評価対象の語彙が前記話題のまとまりに出現する頻度と、該評価対象の語彙が該話題のまとまりを含む大きな話題のまとまりに出現する頻度とを用いて、前記評価結果を得ることを特徴とする請求項６記載の文書要約装置。
【請求項８】　前記抽出手段は、要約作成対象の話題のまとまりから局所的な重要語を抽出し、該要約作成対象の話題のまとまりを含む大きな話題のまとまりから大局的な重要語を抽出し、前記選択手段は、該局所的な重要語と大局的な重要語の両方の出現状況に基づいて、該要約作成対象の話題のまとまりから前記重要文を選択することを特徴とする請求項１記載の文書要約装置。
【請求項９】　前記要約の大きさに応じて、前記話題のまとまりの大きさを決定する決定手段をさらに備えることを特徴とする請求項１記載の文書要約装置。
【請求項１０】　与えられた文書中の処理対象の話題のまとまりに単語が出現する頻度と、該処理対象の話題のまとまりを含む大きな話題のまとまりに該単語が出現する頻度とを用いて、該単語が該処理対象の話題のまとまりに特徴的であるかどうかを評価し、評価結果に基づいて該処理対象の話題のまとまりから重要語を抽出する抽出手段と、
前記重要語の出現状況に応じて要約を生成する生成手段と、
前記要約を出力する出力手段とを備えることを特徴とする文書要約装置。
【請求項１１】　要約作成対象の話題のまとまりから局所的な重要語を抽出し、該要約作成対象の話題のまとまりを含む大きな話題のまとまりから大局的な重要語を抽出する抽出手段と、
前記局所的な重要語と大局的な重要語の両方の出現状況に基づいて、要約を生成する生成手段と、
前記要約を出力する出力手段とを備えることを特徴とする文書要約装置。
【請求項１２】　与えられた文書中の各位置の近傍領域における語彙的結束度を求める手段と、
前記結束度の比較的小さな領域を処理対象から除外して、該結束度の比較的大きな領域を重要部分として抽出する重要箇所特定手段と、
前記重要部分を用いて要約を生成する生成手段と、
前記要約を出力する出力手段とを備えることを特徴とする文書要約装置。
【請求項１３】　コンピュータのためのプログラムを記録した記録媒体であって、
与えられた文書中の話題の階層的構成を認定するステップと、
認定された各話題に関する重要語を抽出するステップと、
前記重要語の出現状況に応じて、各話題のまとまりから重要文を選択するステップと、
前記重要文を用いて要約を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。
【請求項１４】　コンピュータのためのプログラムを記録した記録媒体であって、
与えられた文書中の処理対象の話題のまとまりに単語が出現する頻度を求めるステップと、
前記処理対象の話題のまとまりを含む大きな話題のまとまりに前記単語が出現する頻度を求めるステップと、
得られた２つの頻度を用いて、前記単語が前記処理対象

の話題のまとまりに特徴的であるかどうかを評価するステップと、
評価結果に基づいて、前記処理対象の話題のまとまりから重要語を抽出するステップと、
前記重要語の出現状況に応じて要約を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。
【請求項１５】　コンピュータのためのプログラムを記録した記録媒体であって、
要約作成対象の話題のまとまりから局所的な重要語を抽出するステップと、
前記要約作成対象の話題のまとまりを含む大きな話題のまとまりから大局的な重要語を抽出するステップと、
前記局所的な重要語と大局的な重要語の両方の出現状況に基づいて、要約を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。
【請求項１６】　コンピュータのためのプログラムを記録した記録媒体であって、
与えられた文書中の各位置の近傍領域における語彙的結束度を求めるステップと、
前記結束度の比較的小さな領域を処理対象から除外して、該結束度の比較的大きな領域を重要部分として抽出するステップと、
前記重要部分を用いて要約を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。
【請求項１７】　与えられた文書中の話題の階層的構成を認定し、
認定された各話題に関する重要語を抽出し、
前記重要語の出現状況に応じて、各話題のまとまりから重要文を選択し、
前記重要文を用いて要約を生成することを特徴とする文書要約方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、自然言語などで書かれた機械可読文書の要約を行う装置およびその方法に関し、主として、長めのマニュアルや報告書などの要約（ダイジェスト）を作成し、文書の選別・閲覧のプロセスを支援することを意図している。
【０００２】
【従来の技術】文書を要約するための主要な技術として、文書中の重要語を手掛かりに文を抽出（抜粋）して要約を作成する技術と、文書中の話題のまとまりを認定する技術がある。そこで、これらの従来技術について説明する。
【０００３】まず、要約の作成技術について説明する。従来の文書の要約作成の技術には、大きく分けて２つの方法がある。第１の方法は、文書において重要な部分を認定し、それを抜粋することで要約を作成する方法である。文書の重要な部分は、通常、節、段落、文などの論理要素の単位で抜粋される。以下では、これらを「文」という用語で代表させることにする。
【０００４】第２の方法は、要約として抽出すべき情報の型紙を用意して、その型紙の条件にあった文書中の語句を抽出して要約としたり、その型紙によくあてはまる文を抽出して要約とする方法である。
【０００５】第１の方法は、さらに、何を手掛かりに文の重要性を評価するかによっていくつかの方法に分類される。代表的な方法としては、次の３つが挙げられる。
（１）文書中に出現する単語の頻度と分布を手掛かりとする方法
（２）文と文とのつながり方や文の出現位置を手掛かりとする方法
（３）文の構文的パターンによって重要性を評価する方法
これらのうち、（１）の方法は、まず、文書中に含まれる単語（語句）の重要度を決定し、次に、重要な単語をどれ位含んでいるかによって文の重要度を評価する。そして、評価結果に基づいて重要な文を選択して要約を作成する。
【０００６】単語の重要度を決定する方法としては、文書中の単語の出現頻度（出現度数）そのものを用いる方法、単語の出現度数と一般的な文書集合におけるその単語の出現度数とのずれなどを加味して重みを付ける方法、単語の出現位置に応じて重みを付ける方法などが知られている。単語の出現位置に応じて重みを付ける場合は、例えば、見出しに出現する語を重要とみなすなどの処理が付加される。
【０００７】ここで、対象とする単語は、日本語であれば自立語（特に名詞）のみに、英語であれば内容語のみに限るのが通例である。自立語・内容語とは、実質的な意味を持つ名詞、形容詞、動詞などの語であり、助詞、前置詞、形式名詞など、専ら構文的役割を果たすために使われる語とは区別される。なお、日本語の自立語の形式的定義は、独立した文節を構成できる語というものであるが、ここでは、上述の区別により自立語を定義している。
【０００８】このような要約作成方法には、例えば、次のようなものがある。特開平６－２５９４２４「文書表示装置及び文書要約装置並びにディジタル複写装置」とその発明者による文献（亀田雅之、擬似キーワード相関法による重要キーワードと重要文の抽出、言語処理学会第２回年次大会発表論文集、ｐｐ．９７－１００、１９９６年３月．）では、見出しに含まれる単語を多く含む部分を、見出しに関連の深い重要な部分として抜粋することで要約を作成している。
【０００９】特開平７－３６８９６「文書を要約する方

法および装置」では、文書中に現れる表現（単語など）の複雑さ（語の長さなど）から重要な表現の候補（シード）を選び、重要性の高いシードをより多く含む文を抜粋することで要約を作成している。

【００１０】特開平８－２９７６７７「主題の要約を生成する自動的な方法」では、文書内の単語の出現頻度が大きい順に「主題の用語」を認定し、重要な「主題の用語」を多く含む文を抽出することで要約を作成している。

【００１１】特開平２－２５４５６６「自動抄録生成装置」では、出現頻度の大きい単語を重要語と認定し、重要語が初めて登場する部分や、重要語を多く含む部分、自動的に認定した意味段落の先頭に出現している文などを抜粋することで要約を作成している。

【００１２】次に、文書中の話題のまとまりを認定する方法について説明する。この方法には、大きく分けて次の２つが挙げられる。

（１）文書中で繰り返される語による話題の意味的な結び付き（語彙的結束性：lexical cohesion）に基づく方法

【００１３】この方法（以下、Hearst法と称する）は、意味的に関係の深い部分には、同一の語彙が繰り返し出現するという性質（語彙的結束性）を利用して、話題の切れ目となる部分を自動的に認定するものである。この方法では、まず、文書中の各位置の前後に、段落程度の大きさ（１２０語程度）の窓を設定し、その２つの窓にどれくらい同じ語彙が出現しているかを表す類似度を測定する。類似度としては、次式のような余弦測度（cosine measure）と呼ばれる値が用いられている。

【００１４】

【数１】

$$sim(b_l, b_r) = \frac{\Sigma_t w_{t,b_l} w_{t,b_r}}{\sqrt{\Sigma_t w_{t,b_l}^2 \Sigma_t w_{t,b_r}^2}} \quad (1)$$

【００１５】ここで、b_l と b_r は、それぞれ、左窓（文書の冒頭側の窓）、右窓（文書の末尾側の窓）に含まれる文書の部分を表し、w_{t,b_l}、w_{t,b_r} は、それぞれ、左窓、右窓に出現する単語 t の出現頻度を表す。また、（１）式の右辺の Σ_t は、単語 t に関する総和を表す。

【００１６】（１）式の類似度は、左右の窓に含まれる語彙に共通のものが多いほど大きくなり（最大１）、共通のものがない時に０となる。つまり、この値が大きい部分は、左右の窓で共通の話題を扱っている可能性が高く、逆に、この値が小さい部分は、話題の境界である可能性が高いことになる。

【００１７】Hearst法は、（１）式の値を文書の冒頭から末尾までである間隔（２０語）で測定し、極小となる位置を話題境界とするものである。このとき、類似度の細かい振動を無視するために、次のような調整を行っている。まず、極小点 mp の周囲の部分を切り出す。この部分には、極小点の左側の単調減少している部分と極小点の右側の単調増加している部分が含まれる。

【００１８】次に、切り出された部分の開始点 lp、終了点 rp における類似度 C_{lp}、C_{rp} と、類似度の極小値 C_{mp} との差をもとに、次式の値 ds（depth score）を計算し、これを極小点における類似度の変動量の指標とする。

【００１９】

$$ds = (C_{lp} - C_{mp}) + (C_{rp} - C_{mp}) \quad (2)$$

そして、ds が次式のような閾値 h を越えた極小点だけを話題境界として認定している。

【００２０】

$$h = C_0 - \sigma / 2 \quad (3)$$

ここで、C_0、σ は、それぞれ、文書全体における類似度の平均値と標準偏差である。この方法によれば、類似度がより大きく落ち込んだ部分ほど、話題境界である可能性がより高いとみなされる。また、Hearstは、別法として、繰り返し出現する語の連鎖の開始・終了を手掛かりとして、開始点・終了点の近傍に話題境界を認定する方法も示している。

【００２１】語彙的結束性に基づいて話題のまとまりを認定する方法としては、その他に、日本語の提題助詞「は」のついた文節で始まる文（例えば、「Hearstは、」で始まる文）などを手掛かりとする方法が知られている（特開平７－１６０７１１「書き言葉テキストに対する話題構造認識方法および装置」）。また、その方法とHearstの別法に類似する方法とを併用する方法も知られている（望月源、本田岳夫、奥村学、重回帰分析とクラスタ分析を用いたテキストセグメンテーション、言語処理学会第２回年次大会発表論文集、ｐｐ．３２５－３２８、１９９６年３月．）。

（２）接続詞などで示される文間の連接関係（coherence ralation）から文章構造（rhetorical structure）を求める方法

これらのうち、（１）の語彙的結束性に基づく方法として、まず、Hearstの方法（Marti A. Hearst. Multi-paragraph segmentation of expository text. InProceedings of the 32nd Annual Meeting of Association for ComputationalLinguistics, pp.9-16, 1994.）を簡単に説明する。

【００２２】
【発明が解決しようとする課題】しかしながら、上述した従来の文書要約方法には次のような問題がある。文書における重要語を決定し、重要語を多く含む文を抜粋することで文書の要約を作成する方法では、長めの文書、特に複数の話題に関する文章が混在している複合文書の要約を作成するのが困難である。複数の話題に関する文章が混在している場合、話題毎に重要な単語が異なる可能性が高いので、文書中で出現頻度の大きい単語を単純に重要語とみなすことができない。単純に重要語を決定してしまうと、ある話題に関する重要性を手掛かりに、別の話題の部分から重要でない文が抜粋されてしまうことがあるからである。
【００２３】この問題を解決するためには、文書中の話題のまとまりを認定する必要がある。ここで、語彙的結束性から直接に大きな話題のまとまりを認定する方法がないというもう１つの問題がある。
【００２４】従来の技術では、語彙的結束性に基づいて話題のまとまりを認定する場合、Hearst法のように、数段落程度の大きさあるいは大きくても新聞の１つの記事程度までの大きさのまとまりの認定しか試みられていなかった。そして、それより大きいまとまりは、原文書の章などの書式を手掛かりに認定していた。
【００２５】例えば、前述の特開平２－２５４５６６では、内容的に関連度の高い一連の形式段落（字下げなどにより形式的に区切られた段落）を意味段落として自動認定し、文書全体で出現頻度の大きい語だけでなく、それぞれの意味段落で出現頻度の大きい語も重要語として抽出して、要約を作成している。しかし、この方法では、書式を手掛かりに認定した章や節などの切れ目を、自動的に認定した意味段落の分割点より優先しているため、意味段落は章や節などの切れ目を越えることがなく、より大きな話題のまとまりの認定は試みられていない。
【００２６】話題を認定する方法としても、意味段落認定の主な手掛かりは、形式段落２つ分の範囲で繰り返される語彙であるので、大きな話題のまとまりを認定することは困難である。また、語彙が初めて出現する位置の情報も用いているが、大きな間隔で繰り返される語による結束性などを判定するには十分とは言えない。
【００２７】しかしながら、同じ章に属する節であっても意味的なまとまり方に違いのある場合もあり、このような場合にも的確に大きな話題のまとまりを認定できる方法が必要である。また、文書の書式などは特定の種類の文書に関する約束事であるため、様々な種類の文書の要約を行うためには、文書の種類毎にいちいち経験的な規則を用意しなければならないという問題もある。
【００２８】本発明の課題は、語彙的結束性のような文章一般に見られる現象をもとに、文書中の話題構成を自動的に認定して、話題構成に対応する要約を作成する汎用の文書要約装置およびその方法を提供することである。
【００２９】
【課題を解決するための手段】図１は、本発明の文書要約装置の原理図である。図１の文書要約装置は、構成認定手段１、抽出手段２、選択手段３、および出力手段４を備える。
【００３０】構成認定手段１は、与えられた文書中の話題の階層的構成を認定し、抽出手段２は、認定された各話題に関する重要語を抽出する。また、選択手段３は、重要語の出現状況に応じて、各話題のまとまりから重要文を選択し、重要文を用いて要約を生成する。出力手段４は、生成された要約を出力する。
【００３１】ここで、話題の階層的構成とは、文書を構成する複数の話題のまとまりが２段以上の階層構造を成していることを意味する。この階層的構成は、例えば、文書を構成する複数の大きな話題のまとまりの各々が、１つ以上のより小さな話題のまとまりを含み、小さな話題のまとまりの各々が、１つ以上のさらに小さな話題のまとまりを含むというような話題の包含関係に対応する。
【００３２】構成認定手段１は、例えば、文書全体の大きさの１／４～１／１０程度から段落程度の大きさまで、数種類の大きさの窓幅を設定し、語彙的結束性の強さを表す結束度を各窓幅で測定する。これにより、大きな間隔で繰り返される語などによる大局的な結束性と、小さな間隔で繰り返される語などによる局所的な結束性の両方を捉えることができ、大きな話題のまとまりから小さな話題のまとまりに至る話題の階層的構成を認定することができる。
【００３３】抽出手段２は、例えば、処理対象の話題のまとまりに単語が出現する頻度と、その話題のまとまりを含む大きな話題のまとまりにその単語が出現する頻度とを用いて、その単語が処理対象の話題のまとまりに特徴的であるかどうかを評価する。そして、その評価結果に基づいて、処理対象の話題のまとまりから重要語を抽出する。
【００３４】このように、処理対象の話題のまとまりから重要語を抽出する際に、それを含む上位の話題のまとまりも参照するため、その単語の重要性を上位の話題のまとまりとの関係から評価することができる。このため、話題に関わらず単に多く出現する語を誤って重要語と判定することなく、効率的に重要語を抽出できる。
【００３５】また、抽出手段２は、例えば、要約作成対象の話題のまとまりから局所的な重要語を抽出し、その話題のまとまりを含む大きな話題のまとまりから大局的な重要語を抽出する。そして、選択手段３は、局所的な重要語と大局的な重要語の両方の出現状況に基づいて、要約作成対象の話題のまとまりから重要文を選択し、要約を生成する。

【００３６】このように、要約作成対象の話題のまとまりから重要文を選択する際に、それを含む上位の話題のまとまりに出現する重要語も参照するため、局所的な話題に関する文と大局的な話題に関する文の両方をバランスよく含んだ要約を生成することができる。

【００３７】例えば、図１の構成認定手段１は、後述する図２の話題構成認定部２６に対応し、図１の抽出手段２は図２の重要語抽出部２９に対応し、図１の選択手段３は図２の重要文選択部３０に対応し、図１の出力手段４は図２の出力部３１に対応する。

【００３８】

【発明の実施の形態】以下、図面を参照しながら、本発明の実施の形態を詳細に説明する。図２は、本発明の文書要約装置の基本構成を示している。図２において、文書要約装置（digest generator）１２は、要約対象文書（input document）１１が入力されると、その要約文書１３を作成して出力する。

【００３９】文書要約装置１２は、入力部（input unit）２１、単語認定部（tokenizer）２２、単語辞書（machine readable dictionary）２４、要約粒度決定部２５、話題構成認定部（topic structure detector）２６、重要箇所特定部２８、重要語抽出部（keyword extractor）２９、重要文選択部（sentence selector）３０、および出力部３１を備える。

【００４０】入力部２１は、要約対象文書１１を読み込み、単語認定部２２に渡す。単語認定部２２は、形態素解析部（morphological analyzer）２３を含み、それを用いて要約対象文書１１を言語的に解析して、文書１１に含まれる内容語（名詞・動詞・形容詞・形容動詞など）を切り出す。このとき、形態素解析部２３は、単語辞書２４を参照して、文書１１中の文を、品詞情報付きの単語リストに変換する。単語辞書２４は、形態素解析用の単語辞書であって、単語の表記文字列と品詞・活用の情報との対応関係などを記述している。

【００４１】要約粒度決定部２５は、文書１１の大きさと望ましい要約の大きさから、要約として抽出すべき話題の数を計算し、要約を作成する単位とする話題のまとまりの大きさを決定する。

【００４２】話題構成認定部２６は、話題境界候補区間認定部（topic boundary detector）２７を含み、それを用いて共通の話題について記述している文書の部分（話題のまとまり）を自動認定する。話題境界候補区間認定部２７は、話題構成認定部２６のサブモジュールとして、語彙的結束度の小さい区間を話題境界の候補区間として認定する。語彙的結束度とは、文書１１中の各位置の近傍領域における語彙的結束性の強さを表す指標であり、例えば、各位置の前後に設定したある幅の窓内に出現する語彙の類似性から求められる。

【００４３】重要箇所特定部２８は、語彙的結束度の小さい話題のまとまりを以後の処理の対象から除外し、文書の主要な部分だけが要約に出力されるようにする。重要語抽出部２９は、話題構成認定部２６が認定した話題のまとまりについて、その範囲に出現する語彙が話題に特徴的であるかどうかを評価し、特徴的に出現している語を重要語として抽出する。

【００４４】重要文選択部３０は、それぞれの話題のまとまりについて、重要語を多く含む文を選択し、選択した文を原文書１１での出現順に並べる。そして、必要に応じて選択されなかった文の存在を表す印や段落境界などを挿入することで、要約文書１３を作成する。出力部３１は、作成された要約文書１３を処理結果として出力する。

【００４５】図２の文書要約装置１２によれば、話題構成認定部２６が、共通の話題について記述している文書の部分を話題のまとまりとして認定し、重要語抽出部２９が、それぞれの話題のまとまりに特徴的な語を抽出する。このため、話題の異なる複数の文章が混在している複合文書に対しても、精度よく重要語を抽出することができる。また、重要文選択部３０が、話題のまとまり毎に、そのまとまりに特徴的な重要語を手掛かりに重要文を選択して要約を作成するので、別の話題の重要語の影響で不要な文が抜粋されてしまうこともない。

【００４６】話題構成認定部２６は、語彙的結束度に基づき話題を認定する際に、文書全体の大きさの１／４～１／１０程度の大きい幅の窓により測定した語彙的結束度から、段落程度の大きさの小さい幅の窓により測定した語彙的結束度まで、数種類の語彙的結束度を併用する。このように、大きな間隔で繰り返される語などによる大局的な結束性と、小さな間隔で繰り返される語などによる局所的な結束性の両方に関する情報を利用しているので、話題構成認定部２６は、大きな話題のまとまりから小さな話題のまとまりまで、もれなく話題のまとまりを認定できる。

【００４７】さらに、話題構成認定部２６のサブモジュールである話題境界候補区間認定部２７は、各窓幅の語彙的結束度を移動平均した値を、移動平均の開始点における右結束力および移動平均の終了点における左結束力として扱い、右結束力と左結束力が拮抗している部分（結束力拮抗点）の近傍を、話題境界の候補区間と認定する。

【００４８】移動平均を用いることで、語彙的結束度の小さな変動、すなわち、移動平均区間（移動平均をとる区間）の大きさに比べて小さな範囲内の変動が平滑化される。このため、それぞれの結束力拮抗点の間隔は、ほとんどが移動平均区間の大きさ程度以上となる。これにより、話題構成認定部２６は、それぞれの窓幅の語彙的結束度に基づいて、窓幅程度（移動平均区間の幅程度以上）の大きさの話題のまとまりを選択的に認定できるので、話題の階層的構成を正確に認定することができる。

【００４９】また、重要語抽出部２９は、統計的検定法

により、それぞれの話題のまとまりにおいて有意に多く現れると判定された語を重要語と認定する。このため、話題に関わらず単に多く出現する語を誤って重要語と判定することなく、効率的に重要語を抽出できる。

【００５０】さらに、重要語抽出部２９を使って、要約作成対象の話題のまとまりからだけでなく、要約作成対象の話題のまとまりを含むより大きな話題のまとまりからも大局的な重要語を抽出することができる。これにより、小さな話題のまとまりが並んで、より大きな話題のまとまりを構成しているような場合にも、適切に重要語を抽出することができる。すなわち、個々の小さな話題に特徴的な重要語（副主題を表す語）と、それらに共通する大きな話題に特徴的な重要語（主題を表す語）の両方を区別して抽出することができる。

【００５１】そして、重要文選択部３０は、主題を表す語と副主題を表す語の両方を手掛かりに重要文を選択して要約を作成する。このため、主題と副主題の両方をバランスよく含んだ要約が作成される。

【００５２】また、話題構成認定部２６は、要約粒度決定部２５が決定した大きさ程度の話題のまとまりを認定し、重要語抽出部２９と重要文選択部３０が、この話題のまとまりを単位に要約を作成するので、結果として、抽出すべき話題の数程度で、かつ、同じ程度の大きさの話題を、バランスよく要約に取り込むことができる。

【００５３】さらに、重要箇所特定部２８は、話題構成認定部２６が認定した話題のまとまりのうち、結束度の小さい区間を要約対象から除外する。このため、単に項目を列挙しただけの部分などを抜粋してしまうことがなく、内容の濃い要約を作成することができる。

【００５４】図２の文書要約装置１２は、例えば、図３に示すような情報処理装置（コンピュータ）を用いて構成することができる。図３の情報処理装置は、出力装置４１、入力装置４２、ＣＰＵ（中央処理装置）４３、ネットワーク接続装置４４、媒体駆動装置４５、補助記憶装置４６、およびメモリ（主記憶）４７を備え、それらはバス４８により互いに接続されている。

【００５５】メモリ４７は、例えば、ＲＯＭ（read only memory）、ＲＡＭ（random access memory）などを含み、文書要約処理に用いられるプログラムとデータを格納する。ここでは、図２に示した入力部２１、単語認定部２２、形態素解析部２３、要約粒度決定部２５、話題構成認定部２６、話題境界候補区間認定部２７、重要箇所特定部２８、重要語抽出部２９、重要文選択部３０、および出力部３１が、プログラムモジュールとして格納されている。ＣＰＵ４３は、メモリ４７を利用してプログラムを実行することにより、必要な処理を行う。

【００５６】出力装置４１は、例えば、ディスプレイやプリンタなどであり、ユーザへの問い合わせや要約文書１３などの出力に用いられる。入力装置４２は、例えば、キーボード、ポインティングデバイス、タッチパネルなどであり、ユーザからの指示や要約対象文書１１の入力に用いられる。

【００５７】補助記憶装置４６は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク（magneto-optical disk）装置などであり、要約対象文書１１、要約文書１３、単語辞書２４などの情報を格納する。この補助記憶装置４６に、上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ４７にロードして使用することもできる。

【００５８】媒体駆動装置４５は、可搬記録媒体４９を駆動し、その記録内容にアクセスする。可搬記録媒体４９としては、メモリカード、フロッピーディスク、ＣＤ－ＲＯＭ（compact disk read only memory）、光ディスク、光磁気ディスクなど、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体４９に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ４７にロードして使用することもできる。

【００５９】ネットワーク接続装置４４は、ＬＡＮ（local area network）などの任意のネットワーク（回線）を介して外部の装置と通信し、通信に伴うデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ４７にロードして使用することもできる。

【００６０】図４は、図３の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体４９や外部のデータベース５０に保存されたプログラムとデータは、メモリ４７にロードされる。そして、ＣＰＵ４３は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【００６１】次に、図２の文書要約装置１２の各モジュールの動作を、具体例を用いてより詳細に説明する。要約対象文書としては、（社）電子工業振興協会「自然言語処理システムの動向に関する調査報告書」（平成９年３月）第４章「ネットワークアクセス技術専門委員会活動報告」（ｐｐ．１１７－１９７）を用いた。以下の実施形態では、この文書から文を抜粋してＡ４、１～２枚（１５００文字）程度の要約の作成を試みる。

【００６２】従来、要約の大きさとしては、原文書の１／４程度の大きさが目安とされてきたが、この要約対象文書は８１ページの大きさを持ち、従来の自動要約技術が対象としてきた新聞の社説や記事、数頁程度の論文などに比べて巨大である。また、オンラインで文書を閲覧する場合、画面に一度に表示できるのは２ページ程度が限度である。これらの条件を考慮して、上述のような要約の大きさが決められている。

【００６３】要約対象文書の全体を掲載することは適当ではないので、参考として、要約対象文書中の見出しの一覧を図５から図７に示す。図５は、４．１節および

４．２節の見出しを出現順に示しており、図６は、４．３節の見出しを出現順に示しており、図７は、４．４節の見出しを出現順に示している。

【００６４】図８は、単語認定部２２による単語認定処理のフローチャートである。単語認定部２２は、まず、要約対象文書に形態素解析を施し、品詞付きの単語リストを作成する（ステップＳ１１）。次に、品詞を手掛かりに内容語（名詞・動詞・形容詞・形容動詞）を認定し、内容語に対応する文書の部分に印を付けて（ステップＳ１２）、処理を終了する。図９は、要約対象文書の冒頭部分を示しており、図１０は、単語認定部２２からの対応する出力を示している。

【００６５】図８のステップＳ１１において、形態素解析部２３は、図１１に示すような形態素解析処理を行う。形態素解析部２３は、まず、単語リストをクリアし（ステップＳ２１）、文書の先頭から句点（またはピリオド）などを手掛かりに文の取り出しを試み（ステップＳ２２）、文が取り出せたかどうかを判定する（ステップＳ２３）。

【００６６】文が取り出せれば、次に、単語辞書２４を参照して、文に含まれている単語の候補を求める（ステップＳ２４）。日本語の場合は、図９に示したように、単語と単語の境界が形式的に明示されていないので、文に含まれる部分文字列に対応するすべての単語を候補として求める。例えば、「東京都は大都市だ」という文が取り出された場合、図１２に示すように、この文に含まれるすべての部分文字列が単語の候補となる。

【００６７】これに対して、英語の場合は、単語の境界が空白（スペース）により明示されているため、空白で区切られた文字列に対応する単語について、品詞の候補を求めることが主な処理となる。例えば、“Tokyo is the Japanese capital.”という文が取り出された場合、図１３に示すように、この文に明示的に含まれる５つの単語の基本形と品詞が求められる。

【００６８】次に、形態素解析部２３は、品詞レベルの連接の観点から、妥当な単語の並びを選択し（ステップＳ２５）、選択された単語の並びに品詞と出現位置の情報を付加して、出現順に単語リストに追加する（ステップＳ２６）。次に、次の文の取り出しを試み（ステップＳ２７）、ステップＳ２３以降の処理を繰り返す。そして、ステップＳ２３において文が取り出せなくなると、処理を終了する。

【００６９】図１０の単語認定結果において、墨付き括弧で括られた部分が形態素解析部２３の認定した内容語である。内容語が活用語（動詞・形容詞）の場合、墨付き括弧内で、スラッシュ（／）の前の部分は語幹を表し、スラッシュの後の部分は終止形の活用語尾を表す。これは、後の処理で単語の区別を行うために用いられる情報であるが、この情報の代わりに、品詞と活用を付加しておいてもよい。要するに、例えば、「い／る」と「い／く」のように、語幹だけでは区別の付かない単語を区別するための識別情報であれば、任意のものを用いることができる。

【００７０】また、ステップＳ２５において、単語の並びの妥当性を評価する方法は、形態素解析法として各種のものが知られており、任意のものを用いることができる。例えば、単語の並びの妥当性を訓練データにより推定された出現確率を用いて評価する方法が報告されている（Eugene Charniak. Hidden markov models and two applications. In Statistical Language Learning, chapter 3, pp.37-73. The MIT Press, 1993. ／ Masaaki Nagata. A stochastic japanese morphological analyzer using a forward-DP backward-A*N-best search algorithm. In Proceedings of COLING'94, pp.201-207, 1994.／ 永田昌明、前向きＤＰ後向きＡ*アルゴリズムを用いた確率的日本語形態素解析システム、情処研報ＮＬ－１０１－１０、情報処理学会、１９９４年５月。）。

【００７１】なお、図１０の例では、単語認定部２２がすべての内容語を切り出しているが、切り出しの対象を名詞だけに絞っても構わない。また、英語の文書を対象に処理する場合には、形態素解析処理を行う代わりに、空白で区切られたすべての語のうち、話題に関わらずどこにでも出現する語彙（冠詞、前置詞などの機能語や特に高い頻度で出現する語）を取り除いて、単語を切り出してもよい。このような処理は、単語辞書２４の代わりに、機能語や特に高い頻度で出現する語を格納したストップワードリスト（stop word list）を用意すれば、容易に実現できる。

【００７２】図１４は、要約粒度決定部２５による要約粒度決定処理のフローチャートである。要約粒度決定部２５は、まず、望ましい要約の大きさS_a、望ましい各話題の抜粋量S_t、最小窓幅w_{min}、窓幅比ｒの４つのパラメータをユーザから受け取り（ステップＳ３１）、S_aをS_tで割って抽出すべき話題の概数N_tを求める（ステップＳ３２）。

【００７３】ところで、図１４では、図面の見やすさを考慮して、記号“w_{min}”の添字を、“w_ｍｉｎ”のように下線を付加して記している。他の添字についても、同様の表記法が用いられる場合がある。

【００７４】次に、要約粒度決定部２５は、要約対象文書中の延べ語数W_dを求める（ステップＳ３３）。そして、W_dをN_tで割って抽出すべき話題の大きさの目安w_tを算出した後、初項をw_{min}とし公比をｒとする等比級数の中から、w_tを超えない最大の項を選んで基本窓幅w_1とし（ステップＳ３４）、処理を終了する。このとき、w_1は、次式により計算される。

【００７５】

$$w_1 = w_{min} * (r^{**}int(\log_r((W_d / N_t) / w_{min})))$$

ここで、＊＊はｒをｉｎｔ（）乗することを表し、ｉｎｔ（）は、括弧内の部分の少数点以下を切り捨てて、整数にすることを表している。等比級数の各項は、以降の処理で話題の階層的構成を認定する際に、各層における処理用の窓幅として用いられる。

【００７６】また、別法として、ｉｎｔ（W_d／N_t）の値をそのままw_1として用い、w_{min}の方をw_1＊（１／ｒ）＊＊ｎ（ｎは整数）の形で定義してもよい。さらに、等比級数によらずに、他の任意の方法で窓幅を段階的に縮小していくことも可能である。ただし、後述するように、w_{min}を固定し、２の累乗の値を公比とする等比級数を用いる方法が、計算効率上望ましいことが分かっている。

【００７７】例えば、S_a＝１５００（文字）、S_t＝１５０（文字）、w_{min}＝４０（語）、ｒ＝２、W_d＝１７８１６（語）とすると、抽出すべき話題の数N_tは１０（１５００文字／１５０文字）となる。この場合、話題のまとまりの大きさの目安w_tは１８００語程度（１７８１６語／１０）になるので、これを越えない値として、１２８０語（４０＊2^5）が基本窓幅w_1に採用される。

【００７８】新聞記事などの要約実験によれば、話題の内容を理解可能にするためには、それぞれの話題に関して３文程度（見出し１文＋２～３文：１２０～１５０文字程度）以上の文を抜粋する必要があるという経験的知識が得られている。上記の抜粋量S_tの値は、このような経験的知識に基づき決定されている。また、窓幅w_{min}の値は、新聞記事やレポートなどの平均的な語数により決定されている。

【００７９】次に、話題構成認定部２６の処理について説明する。本実施形態においては、話題のまとまりは前述のHearst法を拡張した方法により認定される。したがって、文書の各位置の語彙的結束度（以下、結束度と略す）を測定し、結束度の小さい部分に話題境界を認定するという方針が採用されている。本実施形態とHearst法の主な違いは、次のような点にある。

（１）結束度を測定するための窓幅の違い

本実施形態では、結束度の計算に用いる窓として、Hearst法より巨大なもの（要約対象文書の全体の語数の１／４～１／１０程度：上述の例では１２８０語）から、段落程度の大きさ（数十語から１００語程度：上述の例では４０語）のものまで、幅の異なるものを数種類併用している。

（２）話題境界の認定手順および認定対象とする話題境界の違い

本実施形態では、Hearst法のように、異なる窓幅で測定したそれぞれの結束度（または類似度）について、結束度が極小となる位置をそのまま話題境界と認定するのではなく、結束度の移動平均（moving average）を用いることで、窓幅程度の大きさのまとまりのみを話題のまとまりとして認定している。

【００８０】これらの違いは、本発明が文書中の話題の階層的構成を認定することに起因する。ここで、話題の階層的構成とは、例えば、ある話題を扱った章の中に関連するいくつかの小さな話題の節が含まれるような、話題の包含関係を有する構成のことである。

【００８１】話題の階層的構成を認定する理由は、小さな話題の部分が並んで、より大きな話題のまとまりを成している場合に、個々の小さな話題に特徴的な重要語（副主題を表す語）と、それらに共通する大きな話題に特徴的な重要語（主題を表す語）の両方を区別して抽出することで、主題と副主題の両方をバランスよく含んだ要約を作成するためである。

【００８２】従来の研究では、数千語レベルの窓幅を使って測定した類似度が、実際の文章における話題の推移と対応するのかどうか、すなわち、実際の文章の話題のまとまりの認定に使えるのかどうかは確かめられていなかった。これを確かめようとした研究がなかったのは、このような単純な測定法で数千語レベルの窓幅を使ってしまうと、測定結果が雑音だらけになって、無意味な変動しか示さないだろうという先入観があったものと推察される。

【００８３】例えば、前述のHearstの文献の結論によれば、シソーラス（類語辞典）などのもっと複雑な情報を使ってより精密な境界認定を実現する可能性が示唆されているが、窓幅については、実験方法の説明の中で簡単に述べられてるだけである。したがって、本実施形態のように、窓幅を極端に変更した場合にどうなるのかなどについての考察は見られない。

【００８４】Hearstは、実験対象毎に微調整して提示した程度の窓幅が、この方法における最適値であり、文章中の副主題に関する数段落程度の大きさの話題のまとまり（passage）を認定するという問題設定が、この方法の限界であろうと考えていた可能性が高い。また、Hearstの目的は、このような数段落程度のまとまりを認定することに限られていたとも考えられる。

【００８５】そこで、Hearstの文献で用いられているものより５～１０倍程度巨大な窓幅によって測定した類似度が、意味のある変動を示すのかどうかを確かめるために、上述の要約対象文書の話題境界をHearst法により認定する実験を行った。この実験により得られた類似度を結束度としてプロットした結果、図１５および図１６のような結束度分布が得られた。

【００８６】これらの図において、横軸の文書における位置は、文書の冒頭から各位置までの間に出現した内容語の延べ数を表す。また、点線は、要約対象文書内の各節の開始位置を表し、長い点線ほど大きい節に対応している。さらに、記号◇でプロットした折れ線グラフは、

（1）式の余弦測度により求めた結束度の系列を表し、記号＊の付いた棒グラフは、結束度の極小点における（2）式のdepth scoreを表し、水平線は（3）式の閾値を表す。

【００８７】結束度の計算においては、図１５では１２８０語幅の窓が用いられ、図１６では６４０語幅の窓が用いられている。また、結束度の系列は、それぞれの窓幅の１／８（１６０語または８０語）の刻み幅で計算され、プロットされている。

【００８８】図１５および図１６を見ると、点線で示された各節の開始位置付近に、閾値を越えるdepth scoreが付与されており、数千語レベルの窓幅を使って測定した結束度も意味のある変動を示すことが分かる。このように、巨大な窓幅による語彙的結束度を用いれば、章・節レベルの話題境界も認定可能である。また、図１５および図１６を比較すると、大きな窓幅の結束度を使うと大きな話題の切れ目が認定でき、小さな窓幅の結束度を使うと小さな話題の切れ目が認定できるという傾向も見てとれる。

【００８９】しかしながら、この実験結果によれば、Hearst法における次のような問題点が指摘される。

（1）大きな窓幅で認定した話題境界と小さな窓幅で認定した話題境界の対応付けが明確でない。

（2）結束度がおおむね単調減少または単調増加している部分の途中に小さな極値が挟まるだけで、depth score は大きく変化してしまうので、これは安定な指標とは言えない。

【００９０】これらの問題点は、例えば、要約対象文書の４．３節の末尾にある参考文献の部分（参）から４．４．１（１）節の部分までに対応する処理結果に現れている。図１５では、この部分は、大局的に見れば、結束度の１つの谷である。この傾向は、図１６でも変わらない。

【００９１】しかし、図１６では、４．３節（参）の部分の幅の狭い小さな山Ｐ１と、４．４．１（２）節の半ばから４．４．１（３）節まであたりの谷Ｐ２とが明確に現れている。このため、６４０語幅の話題境界は、１２８０語幅の話題境界と大きくずれており、このずれは図１５の刻み幅以上に達する。

【００９２】話題の階層的構成を認定する場合、４．４節の開始位置を大きな話題の切れ目と認定し、４．３節（参）の開始位置などはより小さな話題の切れ目と認定したい。しかし、Hearst法のdepth score は安定でないので、これを話題境界に対応する話題の大きさの指標とすることには無理がある。

【００９３】また、depth score が安定でないため、大きな窓幅の結束度により認定された話題境界が、必ずしも小さな窓幅の結束度により認定されるとは限らない。さらに、大きな窓幅の結束度により大きな話題の切れ目だけが話題境界と認定されるわけでもない。このため、Hearst法は、大きな窓幅で大きな話題の切れ目を認定し、小さな窓幅で小さな話題の切れ目を認定するという処理には使えない。

【００９４】本実施形態の話題境界候補区間認定部２７は、これらの問題点を解決するために、移動平均法を応用して、話題境界の区間推定を行う。このとき、各窓幅毎に、結束度を移動平均した値を、移動平均の開始点における右結束力および移動平均の終了点における左結束力として扱い、右結束力と左結束力の拮抗点の近傍を話題境界の候補区間と認定する。

【００９５】移動平均を用いることで、結束度の小さな変動、すなわち、移動平均区間の大きさに比べて小さな範囲内の変動が平滑化される。このため、それぞれの結束力拮抗点の間隔は、ほとんどが移動平均区間の大きさ程度以上となる。

【００９６】これにより、話題構成認定部２６において、次のような話題の階層的構成の認定手順が実現される。

（1）大きな窓幅では大きな話題に対応する話題境界だけを選択的に認定する。

（2）話題境界を区間推定し、大きな窓幅による話題境界と、区間の範囲内で一致しているとみなせるより小さな窓幅による話題境界を求める。そして、両者を同一の話題境界とみなす。

【００９７】図１７は、話題構成認定部２６による話題構成認定処理のフローチャートである。話題構成認定部２６は、まず、基本窓幅w_1、最小窓幅w_{min}、窓幅比ｒの３つのパラメータを要約粒度決定部２５から受け取り（ステップＳ４１）、結束度を測定するための窓幅の集合Wを求める（ステップＳ４２）。窓幅の集合Wは、初項w_0をw_1＊ｒとし、公比を１／ｒとする等比級数から、w_{min}以上の大きさの項を集めて作成される。このとき、Wにおける最大窓幅は、$w_0 = w_1$＊ｒとなる。

【００９８】なお、前述したように、窓幅の集合Wの選び方はこれ以外の方法であってもよいが、計算効率上は、w_{min}を固定し、ｒとして２の累乗の値を用いる方法が望ましい。

【００９９】w_1＝１２８０（語）、w_{min}＝４０（語）、窓幅比ｒ＝２の場合は、最大窓幅w_0は２５６０語（１２８０＊２）となる。次に、話題構成認定部２６は、図１０に示したように、内容語に印が付けられた文書をもとに、文書中の各位置の結束度を、W中のそれぞれの窓幅毎に計算し、結束度系列として記録する（ステップＳ４３）。

【０１００】ここでは、まず、文書の各位置（基準点）の前後に設定した２つの窓の中に出現している語彙（ここでは内容語）を比較し、共通している語彙が多い程大きくなるような値を計算して、その位置における結束度とする。そして、窓の位置を文書の冒頭から末尾に向か

って一定の刻み幅ｔｉｃでずらしながら、結束度の計算を繰り返し、計算した結束度を、文書の冒頭から末尾に向かう系列として記録する。

【０１０１】なお、刻み幅ｔｉｃは、窓幅より小さければいずれの値でも構わないが、例えば、窓幅の１／８というように、窓幅に比例するように設定するのが効率的である。このｔｉｃの値は、ユーザにより指定される。

【０１０２】図１８は、図１０の単語認定結果において設定された２つの窓を示している。ここでは、４０番目の内容語「サービス／する」と４１番目の内容語「内容」の間を基準点として、その前後に窓幅４０語の左窓Ｗ１と右窓Ｗ２が設定されている。この位置における結束度は、次のように計算される。

右結束度＝共通語彙数／左窓中の出現語彙数　（５）
＝６／２９
＝０．２０７

左結束度＝共通語彙数／右窓中の出現語彙数　（６）
＝６／２９
＝０．２０７

結束度＝（右結束度＋左結束度）／２　（７）
＝０．２０７

（５）、（６）、（７）式により得られた各結束度には、次のような意味がある。ある窓に含まれる語がその右側（文書の末尾へ向かう方向）の部分にも出現している場合、その数が多い程、その窓の部分は右側との結び付きが強いと考えられる。この指標が、（５）式の右結束度である。同様に、ある窓とその左側（文書の冒頭へ向かう方向）部分との結び付きの強さを示す指標が、（６）式の左結束度である。そして、基準点においてこれらの２種類の結び付きの強さの指標を平均したものが、（７）式の結束度である。

【０１０６】なお、結束度としては、（７）式の値でなくても、文書中の各位置の近傍領域における語彙的結束性の強さを表す指標として妥当な値であれば、どんなものを用いてもよい。例えば、Ｈearst法のように、左右の窓中の語彙の類似性を表す余弦測度を結束度として用いても構わない。

【０１０７】また、文書中の各位置の近傍領域を２つの窓に分割せずに、その近傍領域に一定回数以上出現している単語の数を結束度とすることもできる。実際、各位置を中心とする近傍領域に、類義語や関連語（例えば、「ウェイター」と「レストラン」）などの意味的に関連する単語が出現する割合に対応する値を、結束度として用いることも報告されている（小嶋秀樹、古郡延治、単語の結束性にもとづいてテキストを場面に分割する試み、電気情報通信学会、信学技報ＮＬＣ９３－７、１９９３年５月．）。

【０１０８】ただし、（７）式に示した結束度の方が、計算が単純であり、解釈もしやすい。以下の説明において、（７）式の結束度を他の結束度と区別する必要がある場合には、「共通語彙比による結束度」と称することにする。

【０１０３】まず、図１９に示すように、左窓Ｗ１と右窓Ｗ２中に出現している内容語の異なり数（窓中の出現語彙数）を数える。図１８では、この出現語彙数は、Ｗ１、Ｗ２ともに２９語ずつである。次に、左窓Ｗ１と右窓Ｗ２の両方に出現している内容語の異なり数（共通語彙数）を数える。図１８では、Ｗ１、Ｗ２中に下線を付けて示した６語が共通語彙となる。

【０１０４】最後に、左窓Ｗ１における共通語彙数と出現語彙数の比を右結束度とし、右窓Ｗ２における共通語彙数と出現語彙数の比を左結束度として、これらの結束度の算術平均を求め、これを基準点における結束度とする。ここでは、次のような結果が得られる。

【０１０５】

【０１０９】次に、図２０は、ステップＳ４３で記録された結束度の系列を示している。ここでは、窓幅ｗの１／４が刻み幅ｔｉｃとして用いられており、文書領域ａ１～ａ１１は、刻み幅ｔｉｃに対応する一定幅の領域である。また、ｃ１は、文書中のａ４とａ５の境界を基準点として計算した、窓幅ｗの結束度を表す。すなわち、ｃ１は、文書領域ａ１～ａ４の部分を左窓の範囲とし、ａ５～ａ８の部分を右窓の範囲として計算された結束度である。

【０１１０】次のｃ２は、窓をｔｉｃ分だけ右へずらして計算された結束度を表し、ａ５とａ６の境界を基準点とする窓幅ｗの結束度である。このようにして、窓をｔｉｃ分ずつ順に右へずらして計算したｃ１，ｃ２，ｃ３，ｃ４，．．．を、文書の冒頭から末尾へ向かう窓幅ｗの結束度系列と呼んでいる。

【０１１１】図２１は、上述の単語認定結果において、文書の冒頭から各基準点までの間に出現した内容語の延べ数を横軸にとり、６４０語の窓幅の結束度系列をプロットしたグラフである。例えば、図２０の結束度ｃ２の場合は、ａ１～ａ５の領域中の内容語の延べ数が、文書における基準点の位置となる。ここでは、６４０語の窓幅の１／８（８０語）を刻み幅ｔｉｃとして、文書の冒頭から末尾に向かって結束度を計算している。

【０１１２】ここで、w^{α} を固定し、窓幅比ｒとして２の累乗の値を用いるのが計算効率上望ましいとした理由について説明する。窓幅比ｒとして２の累乗の値を用いるのが望ましいのは、次の理由による。各窓幅の結束

度の計算においては、文書中のそれぞれの位置で、その位置の前後に設定した２つの窓内の領域とそれらを合わせた領域の３種類の領域に出現する語彙を調べる必要がある。例えば、共通語彙比による結束度を用いる場合には、この３種類の領域に出現する語彙の異なり語数を集計する必要があり、余弦測度による結束度を用いる場合には、この３種類の領域に出現する語彙の出現頻度を集計する必要がある。

【０１１３】図１９では、左窓と右窓の各々の中の語彙数と、これらに共通する共通語彙の数を集計しているが、これらの２つの窓を合わせた領域中の語彙数は、左窓中の語彙数と右窓中の語彙数の和から共通語彙数を差し引いた結果に一致する。したがって、図１９の集計は、上述した３種類の領域に出現する語彙数を集計する処理と同値であり、必要な計算量もほとんど変わらない。

【０１１４】このとき、ｒを２の累乗の値にしておくと、小さな窓幅の結束度の計算のために集計した語彙数（または出現頻度）を、大きい窓幅の結束度の計算でも利用できるようになる。例えば、ｒとして２を用いると、窓幅w_iの結束度の計算において前後の窓を合わせた領域で集計した語彙数（または出現頻度）が、窓幅w_0の結束度の計算における片方の窓内における語彙数（または出現頻度）としても使用できることになる。

【０１１５】また、w_{min}を固定しておくことが望ましいのは次の理由による。w_{min}を固定し、窓幅比ｒとして２の累乗の値を用い、さらに結束度計算の刻み幅ｔｉｃを各窓幅の１／ｎ（ｎは整数）としておくと、要約粒度決定部２５がＷｄを数えるために文書全体を走査する際に、文書をw_{min}／ｎの領域に分割して、結束度系列の計算に便利な形に変換することができる。

【０１１６】例えば、各出現語彙を、ハッシュ表などを用いて、語彙の異なりを識別する語彙番号に変換（数値化）し、w_{min}／ｎ幅の各領域に、出現語彙の語彙番号とその出現頻度を記録しておくことなどが可能になる。こうしておけば、少なくとも結束度系列の計算においては、原文書にアクセスする必要がなくなるので、計算効率が向上する。

【０１１７】また、一般的なＯＳ（オペレーティングシステム）は、原文書の中までアクセスしなくても、原文書の物理的な大きさ（バイト数）を簡単に取得できる機能を持っているのが普通である。

【０１１８】このようなＯＳ上では、最初に、原文書の物理的な大きさで最大窓幅の大体の大きさ（例えば、上限）の見当をつけておき、Ｗｄを数えるために文書全体を走査する際に、結束度系列の計算も同時に行うように工夫することも考えられる。この方法によれば、利用可能な１次メモリの容量が小さい環境でも、原文書へのアクセス回数を減らすことができる。その他にも、計算上の色々な工夫が考えられる。

【０１１９】次に、話題構成認定部２６は、サブモジュールの話題境界候補区間認定部２７を使って、それぞれの窓幅の結束度系列を解析し、結束度の低い区間を話題境界候補区間として認定する（ステップＳ４４）。

【０１２０】図２１に示したように、結束度系列における極小点は、実際の話題境界（点線で示した節の境界）に対応することが多いが、すべての極小点が話題境界に対応するわけではない。話題境界候補区間認定部２７は、結束度系列の極小点を手掛かりに、それぞれの結束度系列の窓幅程度の大きさの話題のまとまりの境界位置を区間推定する。本実施形態では、この処理を、移動平均法を用いて実現している。

【０１２１】次に、話題構成認定部２６は、異なる窓幅の結束度系列に基づいて求めた話題境界候補区間を統合し、大きな窓幅の結束度系列から得られた大きな話題に関する境界と、小さい窓幅の結束度系列からのみ得られる小さい話題に関する境界とを区別して出力する（ステップＳ４５）。これにより、話題構成認定処理が終了する。

【０１２２】ここで、出力される最終的な話題境界は、統合された話題境界候補区間のうち最も小さい窓幅、すなわち最小窓幅の話題境界候補区間を使って認定される。最終的な話題境界の認定に最小窓幅の話題境界候補区間を使う理由は、大きな窓幅の結束度系列は、窓位置の移動に対して鈍感であり、それだけを用いて認定すると、境界位置を十分精密に求めることができないからである。

【０１２３】次に、図１７のステップＳ４４における話題境界候補区間認定処理について、図２０および図２２を使って説明する。ここで用いられる移動平均法は、株価の変動などの統計的分析方法である時系列分析（time series analysis）において、細かい変動を取り除いて大局的な傾向を把握するために使われている。本実施形態では、結束度系列の移動平均値を細かい変動を無視するために用いるだけでなく、それを移動平均の開始点における右結束力および移動平均の終了点における左結束力とみなすことで、話題境界候補区間（低結束度の区間）認定のための直接的な手掛かりとしている。

【０１２４】図２０は、前述したように、結束度の系列ｃ１～ｃ４と文書領域ａ１～ａ１１との関係を示している。結束度系列の移動平均値とは、例えば、（ｃ１＋ｃ２）／２（２項の移動平均）、（ｃ１＋ｃ２＋ｃ３）／３（３項の移動平均）、（ｃ１＋ｃ２＋ｃ３＋ｃ４）／４（４項の移動平均）のように、結束度系列において連続するｎ個の値を算術平均した値である。

【０１２５】図２２は、図２０の結束度系列の移動平均の例と文書領域との関係を示している。ここでは、移動平均の例として、図２０の結束度の２項～４項の移動平均が示され、それぞれの移動平均に関わる結束度の計算において、各文書領域が使用された回数が示されてい

る。このうち、下線を付けた値は、対応する文書領域が移動平均に関わるすべての結束度の計算に用いられていることを表す。

【０１２６】例えば、左上角の値“１”は、ｃ１～ｃ４までの４項の移動平均において、文書領域ａ１が一度だけ左窓の一部として扱われたことを示している。また、その右の値“２”は、ｃ１～ｃ４までの４項の移動平均において、文書領域ａ２が２回左窓の一部として扱われたことを示している。他の使用回数についても、同様である。

【０１２７】結束度は境界の前後の部分の結び付きの強さを表す指標であるので、領域ａ１を左窓に含んで得られた結束度ｃ１を用いて計算された移動平均値も、領域ａ１が右の方向に結び付いているかどうかを示す指標の１つと考えられる。

【０１２８】言い換えれば、移動平均値は、移動平均をとった結束度の左窓部分の領域（ｃ１～ｃ４の４項平均に対してはａ１～ａ７）が右方向に引っ張られる強さ（右結束力）の指標になっていると言える。一方、逆に、移動平均をとった結束度の右窓部分の領域（ｃ１～ｃ４の４項平均に対してａ５～ａ１１）が左方向に引っ張られる強さ（左結束力）の指標になっているとも言える。

【０１２９】ここで、結束力の指標とそれぞれの文書領域との関連性を考察すると、結束度の計算においてより多く窓に含まれていた領域との関連が強いと考えられる。また、語彙的結束性は、一般に、近傍で繰り返される語彙に基づくものほど強いと考えられるので、移動平均をとった結束度の基準点（左右の窓の境界位置）に近い位置にある領域ほど関連が強いとも言える。

【０１３０】例えば、図２２の４項の移動平均については、結束度の基準点は、ａ４とａ５の境界、ａ５とａ６の境界、ａ６とａ７の境界、およびａ７とａ８の境界の４つである。この場合、ａ４は最も多く左窓に含まれており、かつ、これらの基準点に最も近いことが分かる。また、ａ８は最も多く右窓に含まれており、かつ、これらの基準点に最も近いことが分かる。したがって、移動平均値と最も関連の強い領域は、左窓については４、右窓については８となる。

【０１３１】同様にして、３項の移動平均と最も関連の強い領域を選ぶと、左窓についてはａ４、右窓についてはａ７となり、２項の移動平均と最も関連の強い領域を選ぶと、左窓についてはａ４、右窓についてはａ６となる。これらの領域の使用回数は、図２２では斜線を付けて示されている。

【０１３２】以上の考察に基づき、話題境界候補区間設定部２７は、結束度の移動平均値を、移動平均をとった領域内の最初の基準点における右結束力および最後の基準点における左結束力の指標として取り扱う。例えば、ｃ１～ｃ４の４項の移動平均値は、ａ４とａ５の境界における右結束力およびａ７とａ８の境界における左結束力となる。

【０１３３】図２３は、話題境界候補区間設定部２７による話題境界候補区間設定処理のフローチャートである。候補区間設定部２７は、まず、話題構成設定部２６から結束度系列の刻み幅ｔｉｃを受け取り、ユーザから移動平均の項数ｎを受け取る（ステップＳ５１）。

【０１３４】これらのパラメータの値の目安は、刻み幅ｔｉｃについては、例えば、窓幅ｗの１／８～１／１０程度の大きさであり、項数ｎについては、ｗ／ｔｉｃの半分（４～５）程度である。また、移動平均をとる領域の最初の基準点から最後の基準点までの隔たりを、（ｎ－１）＊ｔｉｃにより計算して、それを移動平均の幅ｄ（語）とする。

【０１３５】次に、文書中の各位置ｐについて、ｐ～ｐ＋ｄの範囲内で結束度の移動平均をとり、平均値を位置ｐにおける右結束力として記録する（ステップＳ５２）。この値は、同時に、移動平均をとった範囲の終了位置ｐ＋ｄにおける左結束力としても記録される。

【０１３６】次に、記録された右結束力をもとに、文書中の冒頭から末尾に向かって各位置における右結束力と左結束力の差（右結束力－左結束力）を調べ、その値が負から正に変化する位置を負の結束力指抗点として記録する（ステップＳ５３）。

【０１３７】負の結束力指抗点とは、その位置の左では左結束力が優勢であり、その位置の右では右結束力が優勢であるような点である。したがって、この点の左右の部分は意味的な結び付きが弱いと考えられ、負の結束力指抗点は話題境界の候補位置となる。

【０１３８】次に、認定された結束力指抗点の直前のｄ語以内の範囲で、右結束力が最小となる位置ｍｐを求め、区間［ｍｐ，ｍｐ＋ｄ］を話題境界候補区間と認定して（ステップＳ５３）、処理を終了する。

【０１３９】ここで、左右の結束力の差に基づいて話題境界候補区間を認定する意味を、図２４を使って説明する。図２４は、図２１の５０００語の近傍（４６００語～５４００語付近）における３２０語幅の窓による結束度と左右の結束力の分布を示している。刻み幅ｔｉｃとしては、窓幅の１／８を採用している。

【０１４０】図２４において、記号◇でプロットした折れ線グラフは、結束度Ｃの系列を表し、記号□でプロットした折れ線グラフは、右結束力ＦＣの系列を表し、記号×でプロットした折れ線グラフは、左結束力ＢＣの系列を表す。話題境界候補区間と結束力指抗点を表す２重矩形で示された領域については、後述することにする。

【０１４１】また、点線で示されたｂｐ１、ｂｐ２、ｂｐ３は、左右の結束力の差が０になる３つの点（結束力指抗点）を表す。最初の点ｂｐ１の左側では、左結束力が右結束力より優勢であり、その右側から次の点ｂｐ２までは、右結束力が左結束力より優勢である。さらに、

その右側から最後の点ｂｐ３までは、左結束力が右結束力より優勢であり、その右側では、右結束力が左結束力より優勢である。

【０１４２】したがって、ｂｐ１とｂｐ３は、右結束力と左結束力の差が負から正に変化する負の結束力拮抗点であり、ｂｐ２は、その差が正から負に変化する正の結束力拮抗点である。

【０１４３】このような結束力の変化から、最初の点ｂｐ１の左側の領域は、それより左側のいずれかの部分と比較的強い結束性を示しており、真中の点ｂｐ２の両側の領域は、ｂｐ２に向かって強い結束性を示しており、最後の点ｂｐ３の右側の領域は、それより右側のいずれかの部分と比較的強い結束性を示していることが分かる。実際、左右の結束力と共にプロットした結束度は、ｂｐ１とｂｐ３の近傍で極小値をとり、ｂｐ２の近傍で極大値をとっている。このように、左右の結束力の変化と結束度の変化は密接に関連している。

【０１４４】例えば、図２４の結束力拮抗点ｂｐ３の近傍の曲線で囲まれた部分Ｐ３は、結束度が極小となる部分の１つである。このため、この部分Ｐ３の移動平均（ここでは、４項平均）の値も、Ｐ４およびＰ５における結束力が示しているように、、通常は、極小値をとる。ただし、移動平均をとる領域より狭い範囲で細かい変動がある場合には、移動平均の平滑化作用により、移動平均値すなわち結束力が極小値をとらないこともある。

【０１４５】また、右結束力は移動平均値を移動平均をとる領域の開始位置に記録した指標であるので、右結束力の極小位置は結束度の極小位置の左になる。同様の理由により、左結束力の極小位置は結束度の極小位置の右になる。そして、結束度の変動が十分に大きければ、移動平均をとる領域内に結束力拮抗点が生成されることになる。

【０１４６】また、負の結束力拮抗点の直前のｄ語以内の範囲に右結束力の極小点が存在することは、次のようにして保証される。まず、ある点ｐにおける右結束力、左結束力を、それぞれ、ＦＣ（ｐ）、ＢＣ（ｐ）とおくと、結束力の定義から、

$$ＦＣ（ｐ－ｄ）＝ＢＣ（ｐ）　　（８）$$

であり、拮抗点ｂｐ３では左右の結束力が等しいので、

$$ＦＣ（ｂｐ３－ｄ）（＝ＢＣ（ｂｐ３））＝ＦＣ（ｂｐ３）　　（９）$$

が成り立つ。したがって、拮抗点ｂｐ３の直前の点の右結束力がｂｐ３の値より小さければ、ｂｐ３－ｄからｂｐ３までの範囲、すなわち、ｂｐ３から左にｄ語以内の範囲に、右結束力の極小値が存在することになる。

【０１４７】また、拮抗点ｂｐ３の直前の点の右結束力がｂｐ３の値より小さくない場合は、ｂｐ３の左側において、

$$\begin{aligned}ＦＣ（ｂｐ３－ｄ－１）&＝ＢＣ（ｂｐ３－１）\\&＞ＦＣ（ｂｐ３－１）\\&≧ＦＣ（ｂｐ３）\end{aligned}　　（１０）$$

が成り立つ。さらに、ｂｐ３の右側において、

$$ＦＣ（ｂｐ３）＜ＦＣ（ｂｐ３＋１）　　（１１）$$

または、

$$ＦＣ（ｂｐ３）≧ＦＣ（ｂｐ３＋１）　　（１２）$$

が成り立つ。（１１）式が成り立つとき、（１０）、（１１）より、ｂｐ３－ｄからｂｐ３までの範囲内に、右結束力の極小値が存在することになる。また、（１２）式が成り立つとき、

$$\begin{aligned}ＦＣ（ｂｐ３－ｄ＋１）&＝ＢＣ（ｂｐ３＋１）\\&＜ＦＣ（ｂｐ３＋１）\\&≦ＦＣ（ｂｐ３）\end{aligned}　　（１３）$$

となる。したがって、（１０）、（１３）式より、ｂｐ３－ｄからｂｐ３までの範囲内に、右結束力の極小値が存在することになる。

【０１４８】図２５は、図１７のステップＳ４５において行われる話題境界認定処理のフローチャートである。話題構成認定部２６は、まず、認定された話題境界候補区間を、認定に使った結束度系列の窓幅と、話題境界候補区間内の結束力拮抗点の文書における出現位置とによってソートしてまとめ、話題境界候補区間データの系列Ｂ（ｉ）［ｐ］を作成する（ステップＳ６１）。

【０１４９】ここで、制御変数ｉは、窓幅ｗｉの結束度系列により認定されたことを表す系列番号であり、制御変数ｐは、系列内の各話題境界候補区間を表すデータ番号である。実際には、ｉは、窓幅の大きい順に０，１，２，．．．のような値をとり、ｐは、結束力拮抗点の出現順に１，２，．．．のような値をとる。それぞれのデータＢ（ｉ）［ｐ］は、次のような要素データを含む。

【０１５０】・Ｂ（ｉ）［ｐ］．ｌｅｖｅｌ：話題境界のレベル。初期値はｉ。

・Ｂ（ｉ）［ｐ］．ｒａｎｇｅ：話題境界候補区間。（開始位置、終了位置）の組。

【０１５１】・Ｂ（ｉ）［ｐ］．ｂｐ：結束力拮抗点。（開始位置、終了位置）の組。

ここで、結束力拮抗点は理論的には点であるが、前述の

ように、右結束力と左結束力の差の符号が反転する地点を括抗点として認定しているので、差が負の点を開始位置とし、差が正の点を終了位置とする小さな区間で表される。この区間の幅は、多くの場合、話題境界候補区間の認定に用いられた刻み幅ｔｉｃに一致する。

【０１５２】Ｂ（ｉ）［ｐ］．ｂｐとしては、（開始位置、終了位置）の組を用いる代わりに、別法として、次

$$bp=(DC(rp)*lp-DC(lp)*rp)/(DC(rp)-DC(lp)) \quad (14)$$

次に、話題構成認定部２６は、出力対象とする話題境界のレベルの範囲Ｌを決定する（ステップＳ６２）。出力対象とする話題境界が、基本窓幅w_1、基本窓幅よりひとまわり大きい窓幅（最大窓幅）w_0、および基本窓幅よりひとまわり小さい窓幅w_2の３種類の窓幅によって認定された話題境界候補区間に対応する場合は、Ｌ＝｛０，１，２｝となる。

【０１５４】基本窓幅w_1による話題境界だけでなく、それに準ずる大きさの窓幅w_0、w_2による話題境界も出力対象とするのは、次に行われる重要語抽出処理で話題に特徴的な語彙を選択する際に、これらの話題境界が使われるからである。窓幅比ｒが２で、基本窓幅w_1が１２８０語の場合、w_0＝２５６０（語）、w_1＝１２８０（語）、およびw_2＝６４０（語）の３種類の窓幅の話題境界が出力対象となる。

【０１５５】次に、話題構成認定部２６は、窓幅の異なる話題境界候補区間データを統合する処理を行う。ここでは、１つの系列に属するＢ（ｉ）［ｐ］をまとめてＢ（ｉ）と記し、さらに、次のような表記法を用いて、以下の処理を説明する。

【０１５６】・w_i：Ｂ（ｉ）の系列番号ｉに対応する窓幅。

・d_i：Ｂ（ｉ）に対応する話題境界候補区間の幅（移動平均の幅）。

・ｉｅ：最小窓幅w_{ie}に対応する系列番号。

【０１５７】・｜Ｂ（ｉ）｜：Ｂ（ｉ）におけるデータ番号ｐの最大値。

まず、処理対象を表す系列番号ｉを０に初期化する（ステップＳ６３）。これにより、最大窓幅w_0による話題境界候補区間の系列が処理対象に設定される。次に、処理対象の系列Ｂ（ｉ）に含まれるデータＢ（ｉ）［ｐ］のうち、出力対象外のデータを取り除く（ステップＳ６４）。すなわち、Ｂ（ｉ）［ｐ］．ｌｅｖｅｌ∈Ｌとなるデータ Ｂ（ｉ）［ｐ］だけを残し、その他のデータをＢ（ｉ）から除外する。

【０１５８】そして、ｉをインクリメントしながら、ｉ＋１≦ｉｅである限り、Ｂ（ｉ＋１）を統合対象の系列とする統合処理を行う。この統合処理では、処理対象系列中のそれぞれの話題境界候補区間データＢ（ｉ）［ｐ］（ｐ＝１，．．．，｜Ｂ（ｉ）｜）について、それと同じ付近を境界候補としている統合対象系列中のデータＢ（ｉ＋１）［ｑ］が検出され、両者のデータが統合される。

【０１５９】この処理を途中で打ち切ることも可能であるが、大きい窓幅に対応する系列で処理を打ち切ると境界位置の精密度が落ちることになる。また、この処理にはそれほどの計算量は必要ないので、通常は、最小窓幅に対応する系列まで処理を繰り返す。

【０１６０】具体的な手順は以下の通りである。まず、ｉ＋１とｉｅを比較し（ステップＳ６５）、ｉ＋１≦ｉｅであれば、ｐにｌを代入して（ステップＳ６６）、ｐと｜Ｂ（ｉ）｜を比較する（ステップＳ６７）。ｐ≦｜Ｂ（ｉ）｜であれば、図２６の統合処理を行い（ステップＳ６８）、ｐ＝ｐ＋１として（ステップＳ６９）、ステップＳ６７以降の処理を繰り返す。そして、ステップＳ６７において、ｐが｜Ｂ（ｉ）｜を越えれば、ｉ＝ｉ＋１として（ステップＳ７０）、ステップＳ６４以降の処理を繰り返す。

【０１６１】そして、ステップＳ６５において、ｉ＋１がｉｅを越えれば、統合処理を終了する。ここで、系列Ｂ（ｉｅ）のそれぞれのデータＢ（ｉｅ）［ｐ］について、Ｂ（ｉｅ）［ｐ］．ｒａｎｇｅの区間内で窓幅w_{ie}の結束度が最小となる位置ｍｐを求め、ｍｐとＢ（ｉｅ）［ｐ］．ｌｅｖｅｌとを対応付けて出力する（ステップＳ７１）。これにより、話題境界認定処理が終了する。

【０１６２】次に、図２６の統合処理について説明する。話題構成認定部２６は、まず、統合対象系列中のデータＢ（ｉ＋１）［ｑ］（ｑ＝１，．．．，｜Ｂ（ｉ＋１）｜）の中から、Ｂ（ｉ＋１）［ｑ］．ｂｐ∩Ｂ（ｉ）［ｐ］．ｒａｎｇｅ≠φであり、かつ、Ｂ（ｉ＋１）［ｑ］．ｂｐ≠Ｂ（ｉ）［ｐ］．ｂｐに最も近いデータを、統合対象データとして選択する（ステップＳ８１）。

【０１６３】ここで、Ｂ（ｉ＋１）［ｑ］．ｂｐ∩Ｂ（ｉ）［ｐ］．ｒａｎｇｅ≠φという条件は、Ｂ（ｉ）［ｐ］の話題境界候補区間とＢ（ｉ＋１）［ｑ］の結束力括抗点の区間とが、少なくとも部分的に重複していることを表す。Ｂ（ｉ＋１）［ｑ］．ｂｐが点で指定されている場合は、代わりに、Ｂ（ｉ＋１）［ｑ］．ｂｐ∈Ｂ（ｉ）［ｐ］．ｒａｎｇｅという条件が用いられる。

【０１６４】図２７は、統合対象データの選択例を示し

ている。図２７において、記号◇でプロットした折れ線グラフは、処理対象に対応する６４０語幅の窓による右結束力の系列を表し、記号＋でプロットした折れ線グラフは、６４０語幅の窓による左結束力の系列を表す。また、記号□でプロットした折れ線グラフは、統合対象に対応する３２０語幅の窓による右結束力の系列を表し、記号×でプロットした折れ線グラフは、３２０語幅の窓による左結束力の系列を表す。

【０１６５】また、２重矩形で示された領域のうち、大きな矩形領域が話題境界候補区間に対応し、それに含まれている小さな矩形領域が結束度括抗点に対応する。ここでは、処理対象データＢ（ｉ）［ｐ］の話題境界候補区間と統合対象系列のデータＢ（ｉ＋１）［ｑ］を照合する際に、Ｂ（ｉ）［ｐ］の話題境界候補区間の幅を、飴述の［ｍｐ，ｍｐ＋ｄ］よりｔｉｃ／２だけ左右に拡大し、［ｍｐ－ｔｉｃ／２，ｍｐ＋ｄ＋ｔｉｃ／２］としている。ｔｉｃ／２は、Ｂ（ｉ＋１）［ｑ］に対応する結束度の刻み幅である。

【０１６６】これは、話題境界候補区間の認定精度が結束度系列の刻み幅ｔｉｃに依存するため、ｍｐの本当の位置は（ｍｐ－ｔｉｃ，ｍｐ＋ｔｉｃ）の間と推定されるからである。したがって、処理対象データの話題境界候補区間を広めにとった場合には、（ｍｐ－ｔｉｃ，ｍｐ＋ｄ＋ｔｉｃ）の範囲となる。

【０１６７】ここでは、統合対象データの結束度の刻み幅がｔｉｃ／２であるため、［ｍｐ－ｔｉｃ／２，ｍｐ＋ｄ＋ｔｉｃ／２］を処理対象データの話題境界候補区間としている。このような話題境界候補区間を設定すれば、その幅はｄ＋ｔｉｃ＝ｎ＊ｔｉｃとなる。一方、ｔｉｃ＝ｗ／８、ｎ＝４であるから、話題境界候補区間の幅は、丁度、窓幅の半分ｗ／２となる。

【０１６８】例えば、処理対象データをＢ（２）［６］とすると、その話題境界候補区間Ｂ（２）［６］．ｒａｎｇｅには、統合対象系列の２つのデータの結束力括抗点Ｂ（３）［１１］．ｂｐとＢ（３）［１２］．ｂｐが含まれている。このため、Ｂ（３）［１１］とＢ（３）［１２］が統合対象データの候補となる。これらのうち、Ｂ（３）［１２］．ｂｐの方が、処理対象データの結束力括抗点Ｂ（２）［６］．ｂｐにより近いので、Ｂ（３）［１２］が統合対象データとして選択される。

【０１６９】次に、話題構成認定飴２６は、統合対象データが選択できたかどうかを判定し（ステップＳ８２）、統合対象データが選択できれば、ステップＳ８４の処理を行う。ステップＳ８１において、条件を満たすデータが見つからなかった場合には、処理対象データを認定するときに使った結束度を手掛かりに、擬似的な統合対象データを作成し、Ｂ（ｉ＋１）の系列に挿入する（ステップＳ８３）。そして、ステップＳ８４の処理を行う。

【０１７０】ステップＳ８３では、まず、Ｂ（ｉ）［ｐ］．ｒａｎｇｅの範囲内で、窓幅w_tの結束度が最小となる位置ｍｐを求める。次に、Ｂ（ｉ＋１）［ｑ］．ｂｐ＝［ｍｐ，ｍｐ］、Ｂ（ｉ＋１）［ｑ］．ｒａｎｇｅ＝［ｍｐ－d_{w1}／２，ｍｐ＋d_{w1}／２］と設定して、ｍｐに対応する新たなデータＢ（ｉ＋１）［ｑ］を作成する。

【０１７１】そして、系列Ｂ（ｉ＋１）の中で、Ｂ（ｉ＋１）［ｑ－１］．ｂｐ＜ｍｐかつＢ（ｉ＋１）［ｑ＋１］．ｂｐ＞ｍｐとなるような位置に、作成したデータＢ（ｉ＋１）［ｑ］を挿入する。これにより、疑似的な統合対象データのデータ番号ｑが決定され、それ以降の既存データのデータ番号は書き換えられる。ここで、擬似的な話題境界候補区間データを作成するのは、以降の処理において統合探索範囲を狭め、精密な境界認定を行うためである。

【０１７２】例えば、図２７のＢ（２）［６］を処理対象データとすると、通常の統合対象データの話題境界候補区間Ｂ（３）［１２］．ｒａｎｇｅの幅は、d_3（１６０語）である。このときに、もし、Ｂ（３）［１１］とＢ（３）［１２］のいずれも存在しなかった場合には、図２８に示すように、Ｂ（２）［６］．ｒａｎｇｅの範囲内における窓幅w_2（６４０語）の結束度が最小値をとる位置ｍｐを求める。

【０１７３】そして、その近傍にＢ（３）［１０］．ｒａｎｇｅなどの通常の話題境界候補区間と同じ幅d_3のＢ（３）［ｑ］．ｒａｎｇｅを持つ疑似的なデータＢ（３）［ｑ］を作成する。これにより、ステップＳ８４の処理において、Ｂ（２）［６］．ｒａｎｇｅの幅d_2（３２０語）をd３（１６０語）に絞り込むことができる。

【０１７４】この操作は、処理対象データの話題境界候補区間において、結束度の極小点が明確に１点に決まる場合には、大抵の場合有効である。しかし、話題境界候補区間において結束度にほとんど変動が見られない場合には、その話題境界候補区間を縮小せずに、そのまま用いた方がよいことも飴る。ただし、経験的には、話題境界候補区間において結束度がほとんど変動しないような状況は、あまり多く現れない。

【０１７５】ステップＳ８４では、統合対象データの話題境界レベルＢ（ｉ＋１）［ｑ］．ｌｅｖｅｌを処理対象データの話題境界レベルＢ（ｉ）［ｐ］．ｌｅｖｅｌに変更して、処理対象データＢ（ｉ）［ｐ］と統合対象データＢ（ｉ＋１）［ｑ］の情報を統合する。この処理は、統合対象データＢ（ｉ＋１）［ｑ］の話題境界レベルを小さくすることに対応する。例えば、図２７の統合対象データＢ（３）［１２］の場合は、Ｂ（３）［１２］．ｌｅｖｅｌ＝Ｂ（２）［６］．ｌｅｖｅｌとなる。

【０１７６】これにより、次にステップＳ６４の処理を行うとき、新たに処理対象となる系列Ｂ（ｉ＋１）の中

のデータのうち、少なくとも統合対象データＢ（ｉ＋１）［ｑ］は除外されずに残されることになる。したがって、処理対象データを統合対象データに順次置き換えながら、話題境界候補区間を徐々に絞り込んでいくことができる。

【０１７７】最終的には、統合対象データが系列Ｂ（ｉｅ）から選択され、それぞれの統合対象データＢ（ｉｅ）［ｐ］について、ステップＳ７１の処理が行われる。こうして出力された位置ｍｐが、その話題境界レベルＢ（ｉｅ）［ｐ］．ｌｅｖｅｌにおける話題境界として認定される。

【０１７８】図２９は、こうして得られた話題境界の認定結果を示している。図２９において、２５６０語、１２８０語、６４０語の各窓幅に対応して２重矩形で示された領域のうち、大きな矩形領域が話題境界候補区間に対応し、それに含まれている小さな矩形領域が結束度拮抗点に対応する。Ｂ（０）、Ｂ（１）、Ｂ（２）は、それぞれ、２５６０語、１２８０語、６４０語の各窓幅に対応する系列を表し、２重矩形に添えられた番号［１］，［２］，．．．などは、各系列内のデータ番号を表す。

【０１７９】また、上にある矩形領域ほど大きな窓幅（小さな話題境界レベル）に対応し、下にある矩形領域ほど小さな窓幅（大きな話題境界レベル）に対応する。そして、記号＊の付いた棒グラフは、最終的に求められた話題境界の位置を表す。

【０１８０】後述する重要語抽出処理では、大きな窓幅の結束度に基づいて認定された境界ほど（棒グラフが長いほど）、大きな話題のまとまりに関する境界（話題境界レベルの小さな境界）であるとみなされる。また、小さな窓幅の結束度に基づいて認定された境界ほど（棒グラフが短いほど）、小さな話題のまとまりに関する境界（話題境界レベルの大きな境界）であるとみなされる。

再現率＝（正解数／節境界数）＊１００（％）　　（１５）

適合率＝（正解数／認定境界数）＊１００（％）　　（１６）

ここで、節境界数は、各窓幅における正解データの数を表し、認定境界数は、各窓幅の話題境界レベルに対応する認定境界の数を表し、正解数は、各窓幅において、正解データとの隔たりが４語以内であるような認定境界の数を表す。

【０１８７】例えば、４．４節の先頭の境界は、４．３節（６，０６７語）と４．４節（６，６７０語）の間にあり、小さい方の節の大きさは６，０６７語である。これは最大窓幅の２，５６０語より大きいため、４．４節の先頭の境界は、すべての窓幅において正解データとして扱われる。

【０１８８】また、４．４．１節の先頭の境界は、４．４節の先頭から４．４．１節の先頭までの部分（１１５語）と４．４．１節（２，６４３語）の間にあり、小さい方の節の大きさは１１５語である。したがって、４．４．１節の先頭の境界Ｐ１１より、その前の境界Ｐ１２（４．２．２（３）節の開始位置に対応）の方が、大きな話題の境界であると認定されている。このような若干の食い違いはあるものの、大旨、大きな窓幅によって認定された境界ほど大きな話題の切れ目に対応するという傾向にあることが見てとれる。

【０１８２】また、図３０は、共通語彙比による結束度の代わりに、余弦測度による結束度を用いた場合の話題境界の認定結果を示している。図３０においても、大旨、図２９と同様の傾向が見てとれる。

【０１８３】図３１から図３６までは、各窓幅の結束度を手掛かりに認定された話題境界（認定境界）の特徴を表すデータの集計結果を示している。こううち、図３１から図３３までは、（７）式により求めた共通語彙比による結束度を使った場合の結果を表し、図３４から図３６までは、余弦測度による結束度を使った場合の結果を表す。

【０１８４】図３１と図３４は、本発明の狙い通り、窓幅に応じた大きさの話題のまとまりが認定できているかどうかを調べるために、認定境界の間隔を集計した結果である。これらの集計結果から、窓幅の１～２倍程度の間隔で話題境界が認定されていることが分かる。

【０１８５】また、図３２、３３、３５、３６は、窓幅程度の間隔で認定された境界が、実際に、その大きさ程度の話題のまとまりと対応しているかどうかを調べた結果である。図３２と図３５では、上述の要約対象文書に含まれる節の各境界について、その前後の節の大きさを調べ、小さい方の節の大きさが窓幅以上であるような境界を正解データとして、各窓幅毎に再現率と適合率を集計している。再現率と適合率は、次式により計算された。

【０１８６】

４．１節の先頭の境界は、８０語と４０語の窓幅においてのみ、正解データとして扱われる。

【０１８９】また、図３３と図３６では、要約対象文書に含まれる節の各境界について、その前後の節の大きさを調べ、小さい方の節の大きさが窓幅の１／２以上であるような境界を正解データとして、（１５）、（１６）式により再現率と適合率を集計している。

【０１９０】これらの結果を比較すると、共通語彙比による結果より、余弦測度による結果の方が若干精度が高い。一方、同じ窓幅の結束度に関しては、共通語彙比による結果の方が、認定境界が多めになっている。これは、共通語彙比による結束度の方が、余弦測度による結束度より、繰り返される語彙数の変化に敏感であることによるものと考えられる。

【０１９１】このため、共通語彙比による結束度は、小

さい窓幅では局所的な特異点の影響を受けやすくなり、±４語（合わせて１文程度の大きさ）の精度においては、若干見劣りのする結果を与えている。逆に、大きな窓幅においては、共通語彙比による結束度は、余弦測度による結束度では感知できない変動を拾うことができたものと考えられる。

【０１９２】本発明の実施に当たっては、これらの性質と結束度の計算のためのコストとを考慮し、適切な結束度の計算方法を選択あるいは併用することが望ましい。一般に、共通語彙比による結束度は計算コストが比較的低いため、計算効率を重視する場合にはこれを用いることが推奨される。

【０１９３】次に、結束度と文書中の書式を用いて、話題境界をより精度高く認定する方法を説明する。話題境界候補区間は、図２９に見られるように、実際の節境界を含んでいる確率が高い。このため、図３７に示すような簡単な書式の特徴を手掛かりとして、認定境界の位置を微調整することで、認定結果の精度を上げることが可能である。

【０１９４】図３７には、この調整で用いる書式パターンと境界レベルの関係が示されている。書式パターンとしては、節境界を認定する手掛かりとなる特徴的な文字列が、一般的なＯＳで用いられている正規表現（regular expression）記法により示されている。例えば、「 外１ 」は、「４．１」などのようにピリオドで区切られ

【０１９５】

【外１】

$$\backslash d+\backslash . \backslash d+\backslash . \quad [\ 。]+\$$

【０１９６】た２つの数字で始まり、句点「。」を含まない行を表す。また、境界レベルとしては、上述の話題境界レベルと同様に、小さいほど大きな話題境界に対応するような番号が割り振られている。例えば、４．１節などはレベル１の境界であり、空行（「 外２ 」）はレベル４の境界となる。

【０１９７】

【外２】

$$\$$

【０１９８】図３８は、このような特定の書式パタンを用いた統合処理のフローチャートである。この統合処理は、図２５のステップＳ６８で行われる。図３７のような書式パタンと境界レベルの関係は、あらかじめユーザにより指定されるものとする。

【０１９９】話題構成認定部２６は、まず、与えられた書式パタンを参照しながら、処理対象データの話題境界候補区間Ｂ（ｉ）［ｐ］．ｒａｎｇｅ内を走査し、最も境界レベルが小さく、Ｂ（ｉ）［ｐ］．ｂｐに最も近い節境界の位置ｈｐを求める（ステップＳ９１）。そして、統合対象系列中のデータＢ（ｉ＋１）［ｑ］（ｑ＝１，．．．，｜Ｂ（ｉ＋１）｜）の中から、ｈｐ∈Ｂ（ｉ＋１）［ｑ］．ｒａｎｇｅとなるようなデータＢ（ｉ＋１）［ｑ］を、統合対象データとして選択する（ステップＳ９２）。

【０２００】次に、統合対象データが選択できたかどうかを判定し（ステップＳ９３）、統合対象データが選択できれば、ステップＳ９５の処理を行う。ステップＳ９２において、条件を満たすデータが見つからなかった場合には、節境界ｈｐを用いて擬似的な統合対象データを作成し、Ｂ（ｉ＋１）の系列に挿入する（ステップＳ９４）。そして、ステップＳ９５の処理を行う。

【０２０１】ステップＳ９４では、Ｂ（ｉ＋１）［ｑ］．ｂｐ＝［ｈｐ，ｈｐ］、Ｂ（ｉ＋１）［ｑ］．ｒａｎｇｅ＝［ｈｐ－d_{w1}／２，ｈｐ＋d_{w1}／２］と設定して、ｈｐに対応する新たなデータＢ（ｉ＋１）［ｑ］を作成する。

【０２０２】そして、系列Ｂ（ｉ＋１）の中で、Ｂ（ｉ＋１）［ｑ－１］．ｂｐ＜ｈｐかつＢ（ｉ＋１）［ｑ＋１］．ｂｐ＞ｈｐとなるような位置に、作成したデータＢ（ｉ＋１）［ｑ］を挿入する。これにより、疑似的な統合対象データのデータ番号ｑが決定され、それ以降の既存データのデータ番号は書き換えられる。

【０２０３】ステップＳ９５では、図２６のステップＳ８４と同様に、統合対象データの話題境界レベルＢ（ｉ＋１）［ｑ］．ｌｅｖｅｌを処理対象データの話題境界レベルＢ（ｉ）［ｐ］．ｌｅｖｅｌに変更して、処理対象データＢ（ｉ）［ｐ］と統合対象データＢ（ｉ＋１）［ｑ］の情報を統合する。

【０２０４】このような統合処理を採用した場合は、図２５のステップＳ７１において、話題境の最小位置ｍｐを求める代わりに、図３８のステップＳ９１と同様にして、Ｂ（ｉｅ）［ｐ］．ｒａｎｇｅ内で境界レベルの最も小さい節境界の位置ｈｐを求める。そして、ｈｐとＢ（ｉｅ）［ｐ］．ｌｅｖｅｌを対応付けて出力する。

【０２０５】図３８の統合処理によれば、要約対象文書に含まれる実際の書式パタンを手掛かりとして話題境界が認定されるため、図２６の統合処理に比べて、認定結果の精度が向上する。

【０２０６】次に、重要箇所特定部２８の処理について説明する。重要箇所特定部２８は、話題構成認定部２６が認定した話題境界で区切られた３つのレベルの話題区間のうち、結束度の低いものを以降の要約処理の対象から除外する。

【０２０７】ここで、３つのレベルの話題区間とは、最大窓幅w_0の結束度によって求められた話題境界により区切られた区間と、基本窓幅w_1以上の窓幅の結束度によって求められた話題境界で区切られた区間と、基本窓幅の次に小さい窓幅w_2（＝w_1／ｒ）以上の窓幅の結束度によって求められた話題境界で区切られた区間の３種類の話題区間を指す。低結束度の区間を処理対象から除外するのは、このような区間は、例えば、項目を羅列

したただけの部分のように、内容が薄い部分であることが多いためである。

【０２０８】ここでは、ある話題区間の結束度を、話題の階層的構成における親の話題区間における結束度の平均値と比較することで、その話題区間が低結束度区間であるかどうかを判定する。具体的には、判定対象の話題区間をｂとし、ｂの窓幅をw_bとし、話題区間ｂの中心

$$c < mc + \alpha$$

ここで、αは、低結束度判定の感度を変更するためのパラメータであり、この値が大きいほど、低結束度区間と判定される区間が増える。αとしては、０または親話題区間ａにおけるw_aの標準偏差などを用いることが望ましい。

【０２１０】図３９および図４０は、重要箇所特定部２８による重要箇所特定処理のフローチャートである。重要箇所特定部２８は、まず、文書全体を親話題区間として設定し（図３９、ステップＳ１０１）、最大窓幅w_0により決められた話題区間から、低結束度区間を除外する（ステップＳ１０２）。ここで、文書全体を親話題区間とするのは、w_0に基づく話題境界より上位の話題境界は存在しないためである。

【０２１１】最大窓幅w_0の話題区間は、直接には以後の要約処理の対象とはならないが、w_0の話題区間が除外されると、その中に含まれる基本窓幅w_1の話題区間もすべて除外されるため、要約処理の対象となる話題区間が減少する。

【０２１２】次に、基本窓幅w_1に基づく話題区間から、低結束度区間を除外する。基本窓幅w_1の話題区間の親話題区間は、最大窓幅w_0の話題区間であるので、ステップＳ１０２の処理によって除外されなかったw_0の話題区間を１つずつ取り出し、その中に含まれるw_1の話題区間を除外する。

【０２１３】ここでは、まず、最大窓幅w_0の最初の話題区間を取り出し、それを親話題区間として（ステップＳ１０３）、基本窓幅w_1の話題区間から、低結束度区間を除外する（ステップＳ１０４）。次に、最大窓幅w_0次の話題区間を取り出し、それを親話題区間とする（ステップＳ１０５）。そして、親話題区間が取り出せたかどうかを判定し（ステップＳ１０６）、それが取り出せた場合は、ステップＳ１０４以降の処理を繰り返す。

$$c0 = mc + \alpha$$

次に、親話題区間ａにおいて、窓幅ｗの最初の話題区間を取り出し、それを処理対象話題区間とする（ステップＳ１１３）。そして、処理対象話題区間の中心付近における最大結束度ｃを求める（ステップＳ１１４）。

【０２１９】次に、ｃとｃ０を比較し（ステップＳ１１５）、ｃ＜ｃ０であれば、処理対象話題区間を要約処理の対象から除外する（ステップＳ１１６）。そして、親話題区間ａにおいて、窓幅ｗの次の話題区間を取り出付近における窓幅w_aによる結束度の最大値をｃとし、窓幅w_{a-1}の話題区間のうちｂを含むものを親話題区間ａとし、ａにおける窓幅w_aによる結束度の平均値をｍｃとする。そして、次のような関係が成り立てば、話題区間ｂを低結束度区間であると判定する。

【０２０９】

$$c < mc + \alpha \quad (17)$$

【０２１４】親話題区間が取り出せなかった場合は、基本窓幅w_1の話題区間の除外処理が終了したものとみなし、次に、基本窓幅の次に小さい窓幅w_2の話題区間の除外処理を行う。窓幅w_2の話題区間の親話題区間は窓幅w_1の話題区間であるので、ステップＳ１０４の処理によって除外されなかったw_1の話題区間を１つずつ取り出し、その中に含まれるw_2の話題区間を除外する。

【０２１５】w_2の話題区間を除外することは、要約処理の対象である窓幅w_1の話題区間の中から内容的にまとまりの薄い部分を取り除くことに対応する。これにより、基本窓幅w_1の話題区間の要約文として、余分な内容が抜粋されることを防止できる。

【０２１６】ここでは、まず、基本窓幅w_1の最初の話題区間を取り出し、それを親話題区間として（図４０、ステップＳ１０７）、窓幅w_2の話題区間から、低結束度区間を除外する（ステップＳ１０８）。次に、基本窓幅w_1の次の話題区間を取り出し、それを親話題区間とする（ステップＳ１０９）。そして、親話題区間が取り出せたかどうかを判定し（ステップＳ１１０）、それが取り出せた場合は、ステップＳ１０８以降の処理を繰り返す。

【０２１７】親話題区間が取り出せなかった場合は、窓幅w_2の話題区間の除外処理が終了したものとみなし、処理を終了する。図４１は、図３９のステップＳ１０２、Ｓ１０４、および図４０のステップＳ１０８において呼び出される話題区間除外処理のフローチャートである。話題区間除外処理のサブモジュールは、まず、話題区間の窓幅ｗとその親話題区間ａを呼び出し元から受け取る（ステップＳ１１１）。そして、親話題区間ａにおいて、処理対象の窓幅ｗの結束度の平均値ｍｃを求め、次式により、判定の基準となる基準結束度ｃ０を決定する（ステップＳ１１２）。

【０２１８】

$$c0 = mc + \alpha \quad (18)$$

し、それを処理対象話題区間とする（ステップＳ１１７）。ｃ≧ｃ０であれば、処理対象話題区間を残したまま、ステップＳ１１７の処理を行う。

【０２２０】次に、処理対象話題区間が取り出せたかどうかを判定し（ステップＳ１１８）、それが取り出せた場合は、ステップＳ１１４以降の処理を繰り返す。そして、処理対象話題区間が取り出せなくなれば、処理を終了する。

【００２１】図４２は、図４１のステップＳ１１４において呼び出される最大結束度計算処理のフローチャートである。最大結束度計算処理のサブモジュールは、まず、処理対象話題区間ｂとその話題区間の窓幅ｗを呼び出し元から受け取り（ステップＳ１２１）、話題区間ｂの大きさとｗを比較する（ステップＳ１２２）。

【００２２】話題区間ｂの大きさがｗより大きければ、話題区間ｂから、その両端ｗ／２の部分を除外した区間における最大結束度を求め、その値をｃとして記録して（ステップＳ１２３）、処理を終了する。また、話題区間ｂの大きさがｗ以下であれば、話題区間ｂの中心位置における結束度をｃとして記録し（ステップＳ１２４）、処理を終了する。

【００２３】図４３は、α＝０として、重要箇所特定処理を上述の要約対象文書に適用した結果を示している。図４３において、斜線部分Ｐ２１、Ｐ２２、およびＰ２３は、窓幅w_1（１２８０語）の低結束度区間の除外処理により除外された話題区間を表す。また、横線は、窓幅w_0の各話題区間における窓幅w_1の結束度の平均値ｍｃを表し、矢印は、窓幅w_1の各話題区間の中心付近において、最大結束度ｃに対応する点を表す。

【００２４】例えば、４０００語付近の斜線部分Ｐ２１を見ると、矢印が指す極大値ｃは、明らかに平均値ｍｃより低い値を示しているのが分かる。このため、この話題区間は要約対象から除外されている。他の斜線部分Ｐ２２、Ｐ２３についても同様である。

【００２５】また、ハッチングされた部分Ｐ２４およびＰ２５は、窓幅w_2（６４０語）の低結束度区間の除外処理により除外された話題区間を表す。この処理により除外されなった部分、すなわち、Ｐ２１、Ｐ２２、Ｐ２３、Ｐ２４、およびＰ２５以外の部分は、要約処理の対象となる重要箇所であると認定される。

【００２６】図３９および図４０の重要箇所特定処理では、結束度が閾値より低い話題区間を除外することで重要な話題区間を特定してるが、その代わりに、結束度が閾値以上の話題区間を抽出する処理を行っても、同様の結果が得られる。

【００２７】次に、重要語抽出部２９の処理について説明する。重要語抽出部２９は、話題構成認定部２６が認定し、重要箇所特定部２８が絞り込んだ基本窓幅w_1および最大窓幅w_0の話題区間のそれぞれに特徴的に出現している内容語を選択し、話題区間との対応を付けてそれらを抽出する。

【００２８】ここでは、ある内容語ｔの話題区間ｂにおける出現頻度（出現度数）が期待値を上回り、かつ、次式の対数尤度比Ｌが与えられた閾値（統計的有意水準に対応するχ値）以上であるとき、内容語ｔは話題区間ｂに特徴的であると判定される。

【００２９】

【数２】

$$L=2\left(F_{bt}\log\frac{F_{bt}}{E\left(F_{bt}\right)}+\left(F_{at}-F_{bt}\right)\log\frac{F_{at}-F_{bt}}{F_{at}-E\left(F_{bt}\right)}\right)\qquad(19)$$

【００３０】（１９）式において、F_{bt}は、話題区間ｂにおける単語ｔの出現頻度を表し、F_{at}は、話題区間ｂの親話題区間ａにおける単語ｔの出現頻度を表し、Ｅ（F_{bt}）は、話題区間ｂにおける単語ｔの出現頻度の期待値を表す。Ｅ（F_{bt}）は、親話題区間ａにおける単語ｔの出現密度（出現確率）に、話題区間ｂの大きさを乗じることで得られる。ここで、ある区間における単語の出現密度とは、単語の出現頻度と区間の大きさの比を意味する。

【００３１】（１９）式のＬは、単語ｔの出現確率が話題区間ｂとそれ以外の領域との区別に対して独立であるかどうかに関する尤度比検定の値であり、この値が大きいほど、単語の出現確率がその区別に依存していることを表す。Ｌの自由度νは１であるので、有意水準が１０％なら閾値を６．６３４９０と、有意水準が５％なら閾値を７．８７９４と、有意水準が１％なら閾値を１０．８２７６とすればよい。あるいは、閾値を用いる代わりに、Ｌの大きい順に上位のいくつかの単語を重要語として抽出しても構わない。

【００３２】なお、最大窓幅w_0の話題区間と基本窓幅w_1の話題区間が一致している場合や、窓幅w_0の話題区間のほとんどを１つの基本窓幅w_1の話題区間が占めている場合には、このような検定方法は必ずしもうまく機能しない。このため、ｂの直接の上位の話題区間（例えば、ｂを含む窓幅w_0の話題区間）の大きさがｂの大きさの２倍未満の場合には、親話題区間とした文書全体を用いることにする。

【００３３】図４４および図４５は、重要語抽出部２９による重要語抽出処理のフローチャートである。重要語抽出部２９は、まず、パラメータχ値の統計的有意水準に対応する閾値ｈを、ユーザから受け取る（図４４、ステップＳ１３１）。そして、文書全体を親話題区間候補ａ０とし、ａ０の大きさとａ０に出現しているそれぞれの内容語ｗの出現頻度を求め、それぞれ、Ｓ０とF_{w}として記録する（ステップＳ１３２）。

【００３４】次に、最大窓幅w_0の先頭の話題区間を取り出し、それを親話題区間候補ａ１とする（ステップＳ１３３）。そして、親話題区間候補ａ１の大きさとａ１に出現しているそれぞれの内容語ｗの出現頻度を求め、それぞれ、Ｓ１とF_{w}として記録する（ステップＳ１３

４）。

【０２３５】次に、F_{tw}に記録されたそれぞれの内容語の出現頻度の対数尤度比を、ａ０を基準にして求め、それを閾値ｈと比較して重要語を抽出する（ステップＳ１３５）。そして、ａ１において、基本窓幅ｗ１の最初の話題区間を取り出し、それを重要語抽出対象区間ｂとする（ステップＳ１３６）。

【０２３６】次に、重要語抽出対象区間ｂの大きさと、ｂに出現しているそれぞれの内容語ｗの出現頻度を求め、それぞれ、ＳｂとＦｗとして記録する（図４５、ステップＳ１３７）。そして、Ｓ１と２Ｓｂを比較する（ステップＳ１３８）。ここで、Ｓ１＜２Ｓｂであれば、親話題区間としてａ０を選択し（ステップＳ１３９）、Ｓ１≧２Ｓｂであれば、親話題区間としてａ１を選択して（ステップＳ１４０）、ステップＳ１４１の処理を行う。

【０２３７】ステップＳ１４１では、重要語抽出部２９は、F_{tw}として記録された各内容語の出現頻度の対数尤度比を求め、それを閾値ｈと比較して重要語を抽出する。そして、ａ１において、基本窓幅ｗ１の次の話題区間を取り出し、それを重要語抽出対象区間ｂとする（ステップＳ１４２）。

【０２３８】次に、ｂが取り出せたかどうかを判定し（ステップＳ１４３）、それが取り出せた場合は、ステップＳ１３７以降の処理を繰り返す。そして、ｂが取り

$$E(F_{bt}) = Fat * Sb / Sa$$

ここで、F_{bt}がE（F_{bt}）より大きければ、（１９）式によりｔの尤度比Ｌを求め（ステップＳ１５７）、それを閾値ｈと比較する（ステップＳ１５８）。Ｌがｈ以上であれば、ｔを重要語として抽出する（ステップＳ１５９）。そして、F_{tw}のリストから次の単語を取り出し、それを検定単語ｔとして（ステップＳ１６０）、図４６のステップＳ１５３以降の処理を繰り返す。

【０２４３】ステップＳ１５８においてＬがｈ以上である場合は、話題区間ｂにおける単語ｔの出現頻度が、親話題区間ａにおける出現頻度に比べて特異に大きいものとみなされ、ｔが重要語として抽出される。

【０２４４】ステップＳ１５４、Ｓ１５６、およびＳ１５８で判定結果がＮＯの場合は、ｔを重要語として抽出せずに、ステップＳ１６０以降の処理を行う。そして、ステップＳ１５３においてｔが取り出せなかった場合は、すべての単語の検定が終了したものとみなし、処理を終了する。

【０２４５】図４８は、上述の要約対象文書の見出しのうち、先頭の重要語抽出対象区間（窓幅w_1＝１２８０語の話題区間）に含まれる見出しの例を示しており、図４９は、その区間から抽出された重要語を示している。ここでは、有意水準５％に対応する閾値を用いた。

【０２４６】次に、重要文選択部３０の処理について説明する。重要文選択部３０は、出願人による先願の特願出せなくなると、次に、最大窓幅w_0の次の話題区間を取り出し、それを親話題区間候補ａ１とする（ステップＳ１４４）。

【０２３９】次に、ａ１が取り出せたかどうかを判定し（ステップＳ１４５）、それが取り出せた場合は、図４のステップＳ１３４以降の処理を繰り返す。そして、ａ１が取り出せなくなると、処理を終了する。

【０２４０】図４６および図４７は、図４４のステップＳ１３５および図４５のステップＳ１４１において呼び出される尤度比検定処理のフローチャートである。尤度比検定処理のサブモジュールは、まず、閾値ｈ、親話題区間の大きさＳａ（Ｓ０またはＳ１）、親話題区間における単語の出現頻度F_{aw}（F_{0w}またはF_{1w}）のリスト、検定対象話題区間の大きさＳｂ、および検定対象話題区間における単語の出現頻度F_{bw}のリストを呼び出し元から受け取る（図４６、ステップＳ１５１）。

【０２４１】次に、F_{bw}のリストから最初の単語を取り出し、それを検定単語ｔとする（ステップＳ１５２）。そして、ｔが取り出せたかどうかを判定し（ステップＳ１５３）、それが取り出せた場合は、F_{bt}をｌと比較する（図４７、ステップＳ１５４）。

【０２４２】F_{bt}がｌより大きければ、F_{bt}の期待値（理論値）Ｅ（F_{bt}）を次式により計算して（ステップＳ１５５）、それをF_{bt}と比較する（ステップＳ１５６）。

（２０）

平９－００６７７７「文書要約装置およびその方法」に示された技術を応用して、要約の一部となる重要文を抽出する。

【０２４７】本実施形態における重要文選択処理の特徴は、要約を作成する単位である基本窓幅w_1の話題区間と、話題構成においてそれらの直接の上位に位置する最大窓幅w_0の話題区間の両方に対して、重要語が与えられという点である。このように、階層的に構成された話題区間のそれぞれに重要語を与えて、異なる階層の重要語を併用して重要文を選択するという点において、重要文選択処理は先願の方法とは異なっている。

【０２４８】基本窓幅w_1の話題区間に与えられる重要語は、その語に関連の深い文をその話題区間からのみ抜粋するために用いられる局所的な（ローカルな）重要語である。一方、最大窓幅w_0の話題区間に与えられる重要語は、その下位に位置する複数の要約対象話題区間のそれぞれから関連の深い文を抜粋するために用いられる大局的な重要語である。

【０２４９】特願平９－００６７７７は、少ない抜粋量でも要約文書に重要語を幅広く取り入れることができる方法を示している。この方法によれば、多くの種類の重要語を含む要約文書を生成することができる。これは、１つの文を選択する度に、重要語のリストから、選択された文に含まれる語を取り除いているためである。

【０２５０】このリストは、ユーザの質問文の単語も含んでいるため、注目語リストと呼ばれている。ここで、注目語とは、文書の作成者が書こうとした内容を示すキーワード（端的には、見出しや強調語句）と、ユーザが閲覧したいと思っている文書の事柄を示すキーワード（端的には、文書検索時にユーザが入力する質問文）の両方を含んでいる。

【０２５１】本実施形態においても、数十ページの文書を１ページ程度に要約することを念頭においているため、重要文を選択する度に、注目語リストを更新するという方法を踏襲する。

【０２５２】図５０および図５１は、重要文選択部３０による重要文選択処理のフローチャートである。重要文選択部３０は、まず、最大窓幅ｗ０ の先頭の話題区間を親話題区間ａとして取り出し（図５０、ステップＳ１６１）、ａに対応する重要語を親区間の注目語リストｋｗｌａに登録する（ステップＳ１６２）。そして、ａの先頭部分に見出しが存在するかどうかを判定する（ステップＳ１６３）。見出しの判定には、例えば、図３７に示した書式パタンが用いられる。

【０２５３】見出しが存在する場合には、その見出しに印を付けて必須出力文（必ず要約に含める文）とし、見出しに含まれる内容語を抽出して注目語リストｋｗｌａに加える（ステップＳ１６４）。これにより、見出しに関連する文も、自動的に要約に抜粋されるようになる。

【０２５４】次に、ａに含まれる要約対象の各話題区間をｂｎとし、ｂｎの注目語リストｋｗｌｎを作成する。注目語リストｋｗｌｎには、まず、各話題区間ｂｎに固有の重要語が登録され（ステップＳ１６５）、次に、親話題区間ａの注目語リストｋｗｌａの注目語がマージされる（ステップＳ１６６）。ステップＳ１６３において見出し語が存在しない場合には、そのままステップＳ１６５以降の処理を行う。

【０２５５】次に、ａを親話題区間とするすべてのｂｎを一度に処理して、それぞれの区間から要約に出力する文を１文ずつ選択する（図５１、ステップＳ１６７）。ここで、同じ話題区間ａを親に持つ子話題区間ｂｎを一度にまとめて処理するのは、ａの注目語に関連する文をなるべく多くのｂｎから抜粋することを意図しているためである。このとき、選択された文には、選択済であることを示す印が付けられる。

【０２５６】次に、文が選択できなかった話題区間ｂｎに対応するｋｗｌｎを削除して、その区間に対する選択処理を終了する（ステップＳ１６８）。また、文が選択された話題区間ｂｎについては、選択された文に含まれる注目語を、対応する注目語リストｋｗｌｎから削除する（ステップＳ１６９）。さらに、親話題区間ａの注目語リストｋｗｌａのみに由来する注目語であって、話題区間ｂｎに固有の注目語ではないものが、別の話題区間ｂｘで選択された文に含まれている場合には、その注目語をｂｎの注目語リストｋｗｌｎから削除する（ステップＳ１７０）。

【０２５７】次に、注目語リストｋｗｌｎが残っているかどうか、すなわち、まだ文を選択する余地のある話題区間ｂｎがあるかどうかを判定する（ステップＳ１７１）。そのような注目語リストが残っている場合には、ステップＳ１６７以降の処理を繰り返す。このとき、ステップＳ１６９、Ｓ１７０の処理により空になってしまった注目語リストｋｗｌｎについては、ｂｎに固有の注目語リストとａの注目語リストをマージして、注目語リストｋｗｌｎの初期状態を復元しておく（ステップＳ１７２）。

【０２５８】また、ステップＳ１７１において、注目語リストｋｗｌｎが残っていない場合は、最大窓幅ｗ０ の次の話題区間を親話題区間ａとして取り出す（ステップＳ１７３）。そして、親話題区間ａを取り出せたかどうかを判定し（ステップＳ１７４）、それが取り出せた場合は、図５０のステップＳ１６２以降の処理を繰り返す。

【０２５９】そして、親話題区間が取り出せなかった場合は、ステップＳ１６４で印を付けられた必須出力文とステップＳ１６７で選択された文とをマージし、出現順に並べて要約を作成して（ステップＳ１７５）、処理を終了する。作成された要約に、選択されなかった文の存在を示す印や段落境界などを挿入することで、要約の可読性を高めることも可能である。

【０２６０】ところで、ステップＳ１６７において文を選択できない場合としては、抜粋量の制約により文選択が打ち切られた場合、その時点の注目語リストに含まれている注目語（重要語）を含む文が見つからなかった場合などがある。後者の場合には、もう一度注目語リストの初期状態を復元して文選択を試みることにより、選択可能な文の範囲を広げることができる。

【０２６１】また、ここでは、ユーザが入力した質問文を利用することは示されていないが、例えば、ステップＳ１６２において、質問文から内容語を抽出して注目語リストに加えれば、質問文を容易に処理することができる。

【０２６２】なお、図５０のステップＳ１６５、Ｓ１６６において、ａを親話題区間とするすべての話題区間ｂｎを重要文選択の対象とするのではなく、ｂｎとして重要な話題区間を１つ選び、それだけを対象にして重要文を選択してもよい。この方法は、なるべく短い要約を作成したい場合に有効である。

【０２６３】例えば、非常に短い要約を作成したい場合には、複数の話題区間から重要文を選択すると、それぞれの話題区間について抜粋できる量が、読んで理解できる量を下回ってしまうことがある。そのような場合には、要約作成の対象とする話題を絞って要約することで、その話題については理解可能な量の文を抜粋するこ

とができる。これにより、すべての話題を網羅していても理解するのが困難な要約に比べて、より好ましい要約を作成することができる。

【０２６４】図５２は、このような重要文選択処理のフローチャートである。図５２において、ステップＳ１６１からステップＳ１６４までの処理は、図５０と同様である。重要文選択部３０は、次に、ａを親話題区間とする基本窓幅w_1の話題区間の中から、ａの注目語リストｋｗｌａに含まれる注目語が最も多く出現する区間を選択し、それをｂ０ｎとする（ステップＳ１６５ａ）。そして、ｂ０ｎに固有の重要語を、ｂ０ｎの注目語リストｋｗｌｎに登録する。

【０２６５】次に、親話題区間ａの注目語リストｋｗｌａの重要語を、ｋｗｌｎにマージする（ステップＳ１６５ｂ）。そして、ｂ０ｎを親話題区間とする窓幅w_2の話題区間の中から、ｂ０ｎの注目語リストｋｗｌａに含まれる注目語が最も多く出現する区間を選択し、それを要約対象の話題区間ｂｎとする（ステップＳ１６５ｃ）。こうして、１つの親話題区間ａから１つの要約対象の話題区間ｂｎを選択した後、図５１の処理を行う。

【０２６６】ここで、１つの親話題区間ａからただ１つの要約対象の話題区間ｂｎを選択する代わりに、注目語の出現頻度の大きい順に適当な数の話題区間ｂｎを選択するようにしてもよい。また、要約対象として選択した話題区間ｂｎから十分な量の文を抜粋できない場合には、注目語が次に多く出現する話題区間からも文を選択するようにしてもよい。さらに、ステップＳ１６１およびステップＳ１７３において、ユーザが入力した質問文の内容語に基づき、特定の親話題区間ａのみを処理対象に選ぶことも可能である。

【０２６７】図５３および図５４は、図５１のステップＳ１６７において呼び出される選択処理のフローチャートである。選択処理のサブモジュールは、まず、要約全体の大きさの上限Ｕ１と各話題区間の抜粋量の上限Ｕ２を、ユーザから受け取る（図５３、ステップＳ１８１）。通常、Ｕ１は、前述の望ましい要約の大きさS_0より大きく設定され、Ｕ２は、前述の望ましい話題の抜粋量S_1より大きく設定される。これらのパラメータは、S_0およびS_1をもとにして自動的に算出することもできる。

【０２６８】次に、各話題区間ｂｎ毎に、ｂｎ内に存在する各文と注目語リストｋｗｌｎ内の注目語とを比較し、注目語の出現数（異なり数と延べ数）を、各文毎に記録する（ステップＳ１８２）。そして、Ｕ２を越えない長さの未選択の文の中で、注目語の出現数が最大のものを各話題区間ｂｎから１文ずつ選択する（ステップＳ１８３）。

【０２６９】このとき、ｂｎ内でそれまでに選択済の文があれば、それらの長さの和（ｂｎの抜粋量）と新たに選択する文の長さの合計がＵ２を越えないように、新たな文を選択する。注目語の出現数としては、異なり数と延べ数のいずれか一方を用いてもよく、両方の合計を用いてもよい。そして、選択された文に選択済であることを示す印を付け、ｂｎの抜粋量に選択された文の長さを加算する。

【０２７０】次に、文が選択できなかった話題区間ｂｎに選択終了の印を付け（ステップＳ１８４）、選択済のすべての文の長さの合計ｓを求める（ステップＳ１８５）。そして、ｓをＵ１と比較し（図５４、ステップＳ１８６）、ｓがＵ１以下であれば、処理を終了する。

【０２７１】ｓ＞Ｕ１であれば、すべての話題区間ｂｎに選択終了の印を付け（ステップＳ１８７）、選択された文の中で注目語の出現数が最小のものを除外して、ｓとｂｎの抜粋量をその長さだけ減ずる（ステップＳ１８８）。そして、再びｓをＵ１と比較し（ステップＳ１８９）、まだなおｓ＞Ｕ１であれば、ｓがＵ１以下になるまでステップＳ１８８の処理を繰り返す。

【０２７２】このような選択処理によれば、最終的に出力される要約文書の大きさは、指定された上限Ｕ１以内であることが保証される。上述の要約対象文書の場合、図５５、５６、および５７に示すような要約文書が出力される。ここでは、図面の制約上、１つの要約文書を３つの図に分けて掲載している。この要約文書において、各文の前後に挿入された記号“．．．”は、選択されなかった文の存在を示している。

【０２７３】次に、英語の要約対象文書として、米国出願の明細書の原稿（２３，０００語）を用いた例について説明する。ここでは、次のような処理方法およびパラメータを採用した。

（１）単語認定の方法：ストップワードリストを用いた方法

（２）結束度計算用の窓の幅：

最大窓幅w_0＝２５６０（語）

基本窓幅w_1＝１２８０（語）

窓幅w_2＝６４０（語）

（３）話題境界認定の方法：書式パタンを用いた方法

（４）重要箇所特定処理における低結束度判定用の感度α：

w_0用：$\alpha=-\sigma 0/2$（$\sigma 0$は、窓幅w_0の結束度の標準偏差）

w_1およびw_2用：$\alpha=0$

（５）重要語抽出の閾値：ｈ＝６．６３４９０（有意水準１０％）

（６）重要文選択における抜粋量の上限値：

Ｕ１＝３，０００（bytes）

Ｕ２＝６００（bytes）

要約対象文書の全体を掲載することは適当ではないので、参考として、要約対象文書中の見出しの一覧を図５８に示す。図５８において、（）内の表現は、説明のために付加された見出しの省略形であり、要約対象文書に

は含まれていない。

【０２７４】図５９は、入力された要約対象文書の先頭部分を示しており、図６０は、その部分に対する単語認定処理の結果を示している。図６０において、［］で括られた部分が、認定された単語に対応する。先頭の１文字のみが大文字の単語は、［］では、すべて小文字に置き換えられている。

【０２７５】ここでは、空白および“，”、“．”、“：”、“；”などの区切り記号を手掛かりに単語が切り出され、それらの単語のうち、図６１に示すストップワードリストに含まれる単語が取り除かれた。ストップワードリストとは、重要語として抽出したくない冠詞、前置詞などの単語を、あらかじめ定義したリストである。

【０２７６】また、図６２は、図３８の統合処理において節境界を求めるために用いた書式パタンとその境界レベルを示している。ここでは、先頭が大文字のアルファベットで始まっている行を、境界レベル０の節境界とみなし、最初の空白でない文字が“［”である行を、境界レベル１の節境界とみなしている。

【０２７７】話題境界認定処理においては、話題境界候補区間にこれらの書式パタンに一致する行が見つかった場合は図３８の統合処理を採用し、そうでない場合は図２６の統合処理を採用した。その結果、図６３に示すような認定結果が得られた。

【０２７８】図６３において、節境界の近くに記された（Ｂｇ）、＜１＞などは、図５８に示された見出しの省略形を表す。話題境界候補区間データＢ（ｉ）［ｐ］のうち、＜１＞の節境界Ｐ３１に対応するＢ（０）［１］は、書式パタンを用いなければ、Ｂ（１）［３］と統合されるべきデータである。ここでは、書式パタンを用いた結果として節境界Ｐ３１が検出されている。ところが、Ｂ（１）およびＢ（２）の系列には、Ｐ３１の位置を含むデータが含まれていなかったため、Ｂ（１）［２］およびＢ（２）［３］のような疑似的な統合対象データが生成されている。

【０２７９】図６４は、重要箇所特定処理の結果を示している。図６４において、斜線部分Ｐ４１およびＰ４２は、窓幅w_1（１２８０語）の低結束度区間の除外処理により除外された話題区間を表す。また、横線は、窓幅w_0の各話題区間における窓幅w_1の結束度の平均値を表し、矢印は、窓幅w_1の各話題区間の中心付近において、最大結束度に対応する点を表す。また、ハッチングされた部分Ｐ４３、Ｐ４４、およびＰ４５は、窓幅w_2（６４０語）の低結束度区間の除外処理により除外された話題区間を表す。

【０２８０】なお、窓幅w_0の話題区間に関して低結束度区間の除外処理の感度パラメータαを前述のように調整したのは、要約対象文書の（Ｃｌａｉｍｓ）に対応する節の結束度が他の節に比べて異常に高かったためである。このことは、窓幅w_0の結束度の標準偏差が大きかったことに対応している。実際、窓幅w_0の結束度の平均値が０．４３であるのに対して、その標準偏差は０．１１であった。この認定結果に基づき、図６５、６６、および６７に示すような要約文書が生成された。

【０２８１】以上説明した実施形態においては、日本語および英語の文書を例に挙げて要約処理を説明したが、本発明は、これらの文書以外にも、任意の言語および任意の形式の文書に対して適用され、同様の結果を得ることができる。

【０２８２】また、要約対象文書は、必ずしもディジタル化された電子文書である必要はなく、例えば、紙媒体などに記載された文書でもよい。この場合、イメージスキャナなどの光電変換装置により文書画像を取り込み、文字認識を行うことで、単語認定可能な文書データを作成することができる。

【０２８３】

【発明の効果】本発明によれば、数十頁に渡るような長い文書についても、文書サイズの１／２～１／４程度の大きな話題のまとまりから、段落程度の大きさ（数十語から１００語程度）の話題のまとまりまで、任意の大きさの話題のまとまりの階層的構成を、語彙的結束性という文章一般に見られる現象に基づいて認定することができる。

【０２８４】さらに、それぞれの話題のまとまりから適切な内容を抜粋して、話題の階層的構成に対応する要約を作成することができる。これにより、従来は取り扱いが難しかった、複数の話題に関する文章が混在した複合文書の要約が可能になる。

【０２８５】また、要約作成の単位とする話題のまとまりの大きさを自動的に決定し、要約対象を重要な話題のまとまりに絞り込むことで、要約として出力すべき大きさに応じて、適切な粒度の話題をバランスよく取り込んだ要約を作成することが可能になる。

【図面の簡単な説明】

【図１】本発明の文書要約装置の原理図である。

【図２】本発明の文書要約装置の構成図である。

【図３】情報処理装置の構成図である。

【図４】記録媒体を示す図である。

【図５】第１の要約対象文書中の見出しを示す図（その１）である。

【図６】第１の要約対象文書中の見出しを示す図（その２）である。

【図７】第１の要約対象文書中の見出しを示す図（その３）である。

【図８】単語認定処理のフローチャートである。

【図９】第１の入力文書を示す図である。

【図１０】第１の単語認定結果を示す図である。

【図１１】形態素解析処理のフローチャートである。

【図１２】日本語の辞書引きの例を示す図である。

【図１３】英語の辞書引きの例を示す図である。
【図１４】要約粒度決定処理のフローチャートである。
【図１５】第１の結束度分布を示す図である。
【図１６】第２の結束度分布を示す図である。
【図１７】話題構成認定処理のフローチャートである。
【図１８】左窓と右窓を示す図である。
【図１９】窓内の語彙数を示す図である。
【図２０】結束度の系列を示す図である。
【図２１】第３の結束度分布を示す図である。
【図２２】移動平均と文書領域の関係を示す図である。
【図２３】話題境界候補区間認定処理のフローチャートである。
【図２４】結束力分布を示す図である。
【図２５】話題境界認定処理のフローチャートである。
【図２６】第１の統合処理のフローチャートである。
【図２７】統合対象データを示す図である。
【図２８】疑似データの作成方法を示す図である。
【図２９】話題構成の第１の認定結果を示す図である。
【図３０】話題構成の第２の認定結果を示す図である。
【図３１】第１の話題境界の間隔を示す図である。
【図３２】第１の再現率と適合率を示す図である。
【図３３】第２の再現率と適合率を示す図である。
【図３４】第２の話題境界の間隔を示す図である。
【図３５】第３の再現率と適合率を示す図である。
【図３６】第４の再現率と適合率を示す図である。
【図３７】第１の書式パタンと境界レベルを示す図である。
【図３８】第２の統合処理のフローチャートである。
【図３９】重要箇所特定処理のフローチャート（その１）である。
【図４０】重要箇所特定処理のフローチャート（その２）である。
【図４１】話題区間除外処理のフローチャートである。
【図４２】最大結束度計算処理のフローチャートである。
【図４３】重要箇所の第１の特定結果を示す図である。
【図４４】重要語抽出処理のフローチャート（その１）である。
【図４５】重要語抽出処理のフローチャート（その２）である。
【図４６】尤度比検定処理のフローチャート（その１）である。
【図４７】尤度比検定処理のフローチャート（その２）である。
【図４８】話題区間中に含まれている見出しを示す図である。
【図４９】話題区間から抽出された重要語を示す図である。
【図５０】重要文選択処理のフローチャート（その１）である。
【図５１】重要文選択処理のフローチャート（その２）である。
【図５２】重要文選択処理の他のフローチャートである。
【図５３】選択処理のフローチャート（その１）である。
【図５４】選択処理のフローチャート（その２）である。
【図５５】第１の要約結果を示す図（その１）である。
【図５６】第１の要約結果を示す図（その２）である。
【図５７】第１の要約結果を示す図（その３）である。
【図５８】第２の要約対象文書中の見出しを示す図である。
【図５９】第２の入力文書を示す図である。
【図６０】第２の単語認定結果を示す図である。
【図６１】ストップワードを示す図である。
【図６２】第２の書式パタンと境界レベルを示す図である。
【図６３】話題構成の第３の認定結果を示す図である。
【図６４】重要箇所の第２の特定結果を示す図である。
【図６５】第２の要約結果を示す図（その１）である。
【図６６】第２の要約結果を示す図（その２）である。
【図６７】第２の要約結果を示す図（その３）である。

【符号の説明】

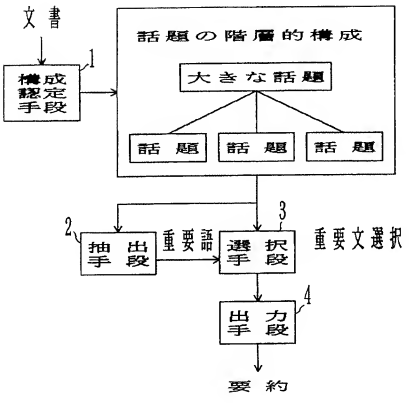
１　構成認定手段
２　抽出手段
３　選択手段
４　出力手段
１１　要約対象文書
１２　文書要約装置
１３　要約文書
２１　入力部
２２　単語認定部
２３　形態素解析部
２４　単語辞書
２５　要約粒度決定部
２６　話題構成認定部
２７　話題境界候補区間認定部
２８　重要箇所特定部
２９　重要語抽出部
３０　重要文選択部
３１　出力部
４１　出力装置
４２　入力装置
４３　ＣＰＵ
４４　ネットワーク接続装置
４５　媒体駆動装置
４６　補助記憶装置
４７　メモリ
４８　バス

49　　50

４９　可搬記録媒体　　５０　データベース

【図１】

本　発　明　の　原　理　図



【図１２】

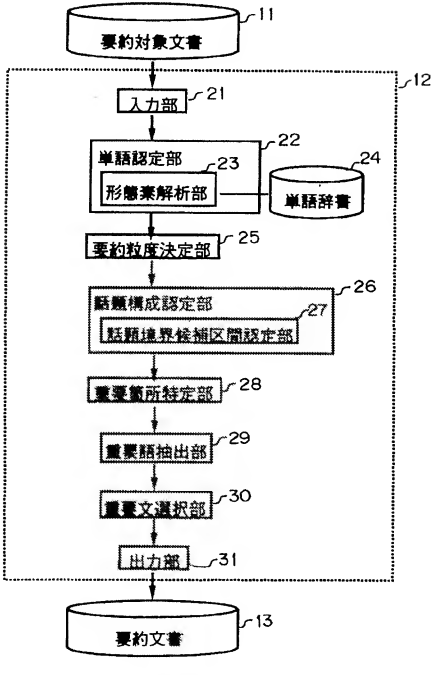
日本語の辞書引きの例を示す図

入力文　東京都は大都市だ

	見出し（語幹）	品詞
候補単語	東	名詞
	東京都	名詞
	京都	名詞
	は	助詞「は」
	大	接頭語
	都市	名詞
	だ	助動詞「だ」

【図２】

文書要約装置の構成図



【図５８】

第2の要約対象文書中の
見出しを示す図

TITLE: SGML Type Document Managing Apparatus and Managin Method
Background of the Invention (Bg)
Field of the Invention
Description of the Related Art
Summary of the Invention
Brief Description of Drawings
Description of Preferred Embodiment (Preferred Embodiment)
[1] Partial editing DTD (<1>)
[2] Revision history information (<2>)
What is claimed is: (Claims)

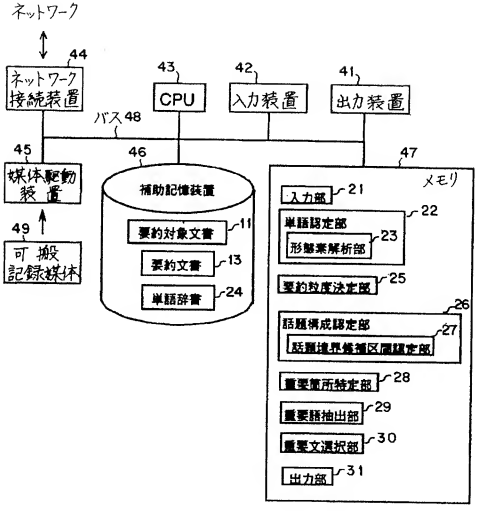
【図６２】

第2の書式パタンと
境界レベルを示す図

書式パタン	境界レベル
^[A-Z]	0
^□*\[	1

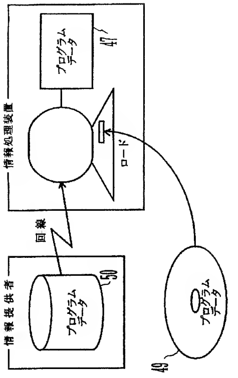
【図３】

情報処理装置の構成図



【図4】

記録媒体を示す図



【図9】

第1の入力文書を示す図

4.　ネットワークアクセス技術委員会

4.1　調査の概要

インターネットは予想されていた以上の早さで急速に普及している。業務はもちろん特に家庭での利用が急速に広がっている。それにともなってインターネットを通じて提供される情報も多種多様化している。そのため、インターネットを利用する上での社会的・技術的なさまざまな要請が顕在化し、それらへの対応が急務となっている。これらの要請の中には、インターネットの健全な運営に関わる認証などの問題とともに、インターネットのサービス内容を高度化するや使い勝手をよくすることに関わる知的情報アクセスの問題が重要な問題として認識されている。

本委員会は、特に、ネットワークアクセスにおける知的情報アクセスの問題について、広範かつ詳細な専門的な立場から調査・研究することを目的としている。その調査・研究では、特に、自然言語処理の分野における技術的な側面に着目し、自然言語処理に関する技術の現状を整理しかつ将来の発展を見通すことにより、今後において協力してあるいは個別で重点的に取り組むべき課題を明らかにすることを目的としている。

【図31】

第1の話題境界の間隔を示す図

		話題境界の間隔（語）		
窓幅	認定境界数	平均	最大値	最低値
2,560	4	3,563	6,145	1,310
1,280	9	1,782	3,010	851
640	17	990	2,185	375
320	38	457	1,000	205
160	74	238	475	40
80	150	111	245	35
40	322	55	165	15

【図５】

第1の要約対象文書中の見出しを示す図(その1)

【図6】

第1の要約対象文書中の見出しを示す図(その2)

４．３　ネットワーク上の検索サービス
４．３．１　検索サービスの調査
（１）WWW検索サービスの概要
（２）情報収集／検索方式
（３）情報提示方式
（４）今後の課題
４．３．２　検索技術の動向
（１）キーワード抽出
（２）文書自動分類
（３）要約・抄録技術
（４）分散検索
４．３．３　電子出版及び電子図書館
（１）電子出版
（２）電子図書館

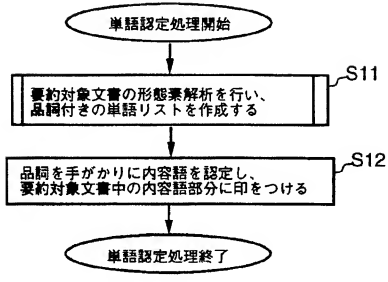
【図7】

第1の要約対象文書中の見出しを示す図(その3)

4．4．検索エンジン
4．4．1．日本語の全文検索技術の動向
（1）文字列検索アルゴリズム
（2）インデックス作成法
（3）日本語の全文検索技術
（4）製品化動向
（5）今後の課題
4．4．2．有限オートマトンによる自然言語処理技術の動向
（1）有限変換器のコンパクト化
（2）文字列パタン照合
（3）書き換え規則,Two-level モデル
（4）形態素解析, 構文解析
（5）まとめ
4．4．3　情報フィルタリング技術の動向
（1）内容に基づくフィルタリング (content-based filtering)
（2）協調フィルタリング (collaborative filtering)
（3）ユーザモデリング
（4）まとめ
4．4．4 情報抽出 / 統合技術の動向
（1）検索ナビゲーション技術
（2）情報統合技術
（3）情報の可視化技術

【図８】

単語認定処理のフローチャート



【図10】

第10単語認定結果を示す図

4.1 【調査/する】の【概要/】

【インターネット/】は【予想/する】されて【い/る】た以上の【早さ/】で【急速/】に【普及/する】して【い/る】る。【業務/】はもちろん各に【家庭/】での【利用/する】が【急速/】に【広が/る】って【い/る】る。それに【とも/う】って【インターネット/】を通じて【提供/する】される【情報/】も【多種多様化/】して【い/る】る。そのため、【インターネット/】を【利用/する】する上での【社会的/】・【技術的/】なさまざまな【課題/する】が【顕在化/する】し、それらへの【対応/する】が【急務/】と【な/る】って【い/る】る。これらの【実際/する】の中には、【インターネット/】の【概念/】な【運営/する】に【関わ/る】る【設置/する】などの【問題/】とともに、【インターネット/】の【サービス/する】【内容/】を【高度化/する】するや【使/う】い【勝手/】をよく【する/】ことに【関わ/る】る【知的/】【情報/】【アクセス/する】の【問題/】が【重要/】な【問題/】として【認識/する】されて【い/る】る。

本【委員会/】は、特に、【ネットワーク/】【アクセス/する】における【知的/】【情報/】【アクセス/する】の【問題/】について、【広範/】かつ【詳細/】な【専門/】的な【立場/】から【調査/する】・【研究/する】することを【目的/】として【い/る】る。その【調査/する】・【研究/する】では、特に、【自然言語/】【処理/する】の【分野/】における【技術的/】な【側面/】に【着目/する】し、【自然言語/】【処理/する】に関する【技術/】の【現状/】を【整理/する】しかつ将来の【発展/する】を【見通/す】すことにより、今後において【協力/する】してあるいは【個別/】で【重点的/】に【取り組/む】むべき【課題/】を【明らか/】にすることを【目的/】として【い/る】る。

【図32】

第1の再現率と適合率を示す図

窓 幅	節境界数	一 致	再現率	適合率
2,560	2	1	50.0 %	25.0 %
1,280	3	2	66.7 %	22.2 %
640	12	5	41.7 %	29.4 %
320	18	6	33.3 %	15.8 %
160	33	11	33.3 %	14.9 %
80	43	15	34.9 %	9.4 %
40	45	18	40.0 %	5.6 %

【図33】

第2の再現率と適合率を示す図

窓 幅	節境界数	一 致	再現率	適合率
2,560	9	1	33.3 %	25.0 %
1,280	12	4	33.3 %	44.4 %
640	18	6	33.3 %	85.3 %
320	38	8	34.2 %	21.1 %
160	43	13	30.2 %	17.8 %
80	45	15	33.3 %	9.4 %
40	46	19	41.3 %	5.8 %

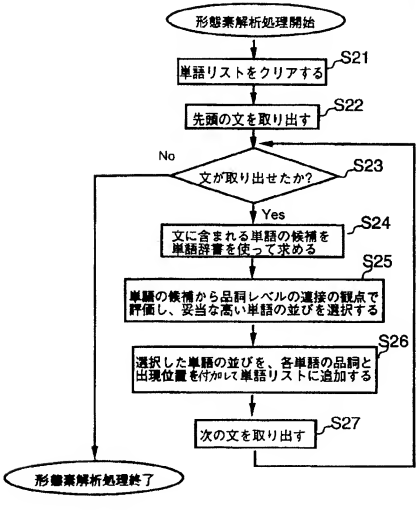
【図34】

第2の話題境界の間隔を示す図

窓 幅	認定境界数	話題境界の間隔（語）			
		平 均	最 大 値	最 低 値	
2,560	3	4,454	6,370	2,535	
1,280	10	1,620	2,535	805	
640	17	990	1,345	605	
320	30	575	1,100	170	
160	70	251	470	85	
80	147	120	290	40	
40	308	58	165	10	

【図１１】

形態素解析処理のフローチャート



【図３５】

第３の再現率と適合率を示す図

窓　幅	簡略界数	一　致	再現率	適合率
2,560	2	1	50.3 %	33.3 %
1,280	3	2	66.7 %	20.0 %
640	12	5	41.7 %	26.4 %
320	18	7	38.9 %	23.3 %
160	33	13	39.4 %	18.6 %
80	43	18	41.9 %	12.2 %
40	45	21	46.7 %	6.8 %

【図13】

英語の辞書引きの例を示す図

入力文　Tokyo is the Japanese capital.

	headword	base(root) form	part of speech
候補単語	Tokyo	Tokyo	proper noun
	is	be	be verb (the third person singular present form)
	the	the	definite article
	Japanese	Japanese	proper noun
	Japanese	Japanese	adjective
	capital	capital	noun

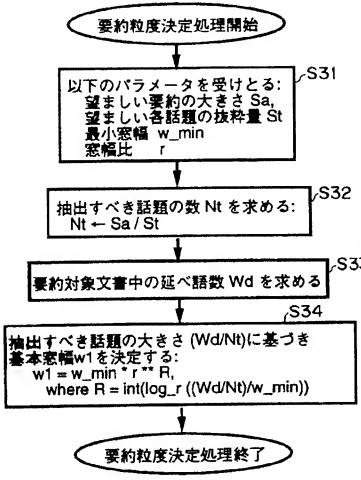
【図19】

窓内の語彙数を示す図

窓幅	40語
精実度測定位置	40(語)
左窓(W1)中の語彙数(異なり語数)	29語
右窓(W2)中の語彙数(異なり語数)	29語
共通語彙数(異なり語数)	6語 (いる，問題，調査する，情報，関わる，技術的)

【図１４】

要約粒度決定処理のフローチャート



【図３６】

第４の再現率と適合率を示す図

窓 幅	箱境界数	一 致	再現率	適合率
2,560	3	1	33.3 %	33.3 %
1,280	12	4	33.3 %	40.0 %
640	18	7	38.9 %	41.2 %
320	38	10	30.3 %	33.3 %
160	43	14	32.6 %	20.0 %
80	45	18	40.0 %	12.2 %
40	46	21	45.7 %	6.8 %

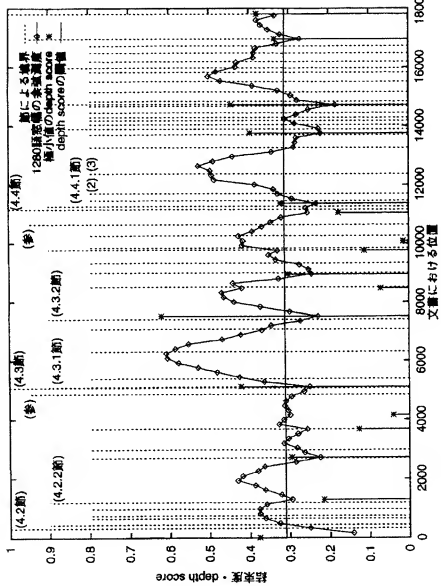
【図３７】

第1の
書式パターンと境界レベルを示す図

書式パターン	境界レベル
^\d+[^.]+$	0
^\d+\.\d+\.[^.]+$	1
^\d+\.\d+\.\d+\.[^.]+$	2
^(\d+)[^.]+$	3
^.$	4

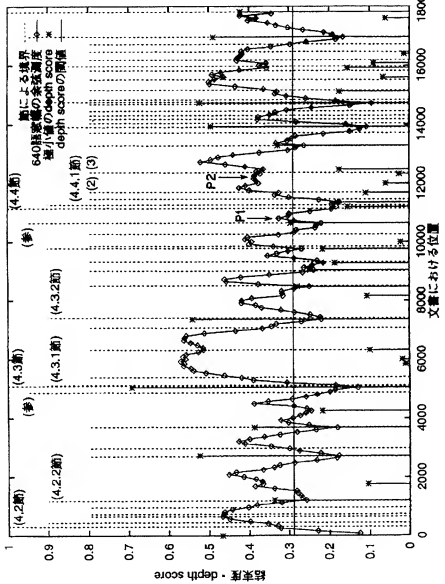
【図１５】

第1の結束度分布を示す図



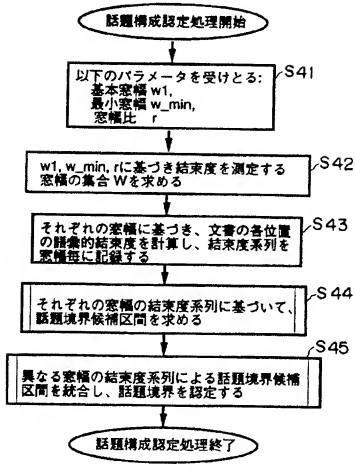
【図１６】

第2の結束度分布を示す図



【図１７】

話題構成認定処理のフローチャート



【図18】

左窓と右窓を示す図

【図20】

結束度の系列を示す図

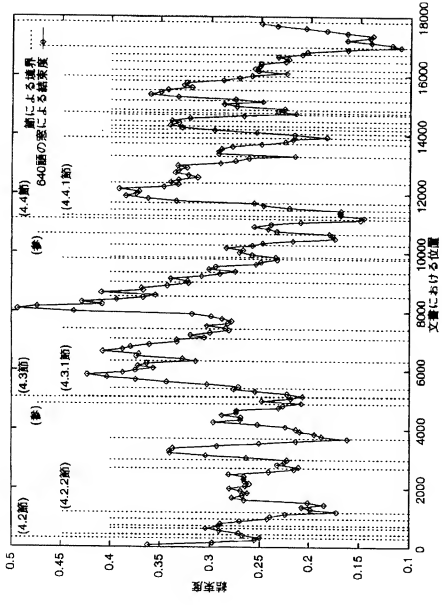
【図22】

移動平均と文書領域の関係を示す図

項目数		文書領域の使用回数										
		a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11
4項平均	左窓	1	2	3	4	3	2	1	0	0	0	0
(c1～c4)	右窓	0	0	0	0	1	2	3	4	3	2	1
3項平均	左窓	1	2	3	3	2	1	0	0	0	0	
(c1～c3)	右窓	0	0	0	0	1	2	3	3	2	1	
2項平均	左窓	1	2	2	2	1	0	0	0	0		
(c1, c2)	右窓	0	0	0	0	1	2	2	2	1		

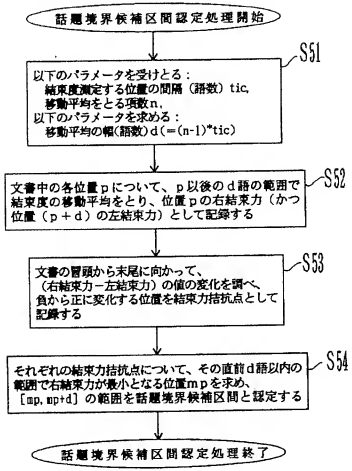
【図２１】

第３の結束度分布を示す図



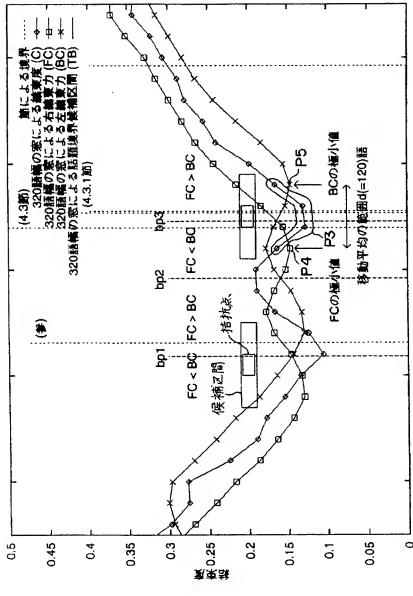
【図２３】

話題境界候補区間認定処理のフローチャート



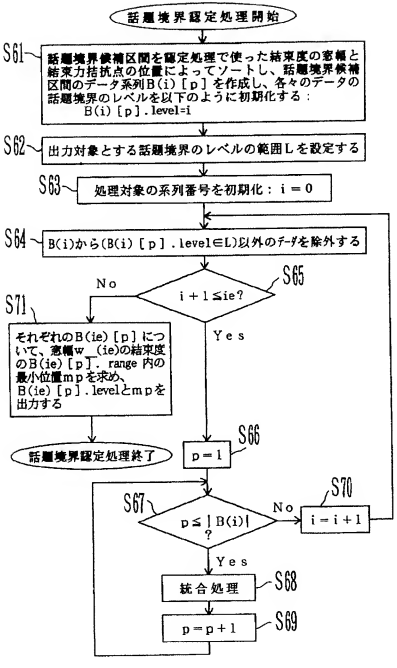
【図２４】

結束力分布を示す図



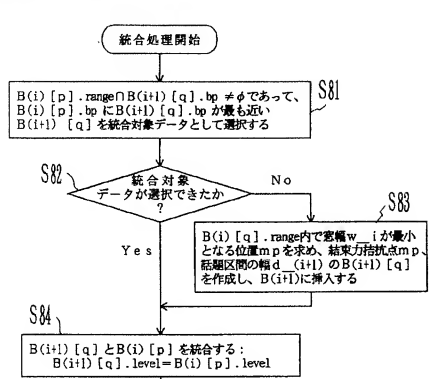
【図２５】

話題境界認定処理のフローチャート



【図２６】

第１の統合処理のフローチャート



【図４８】

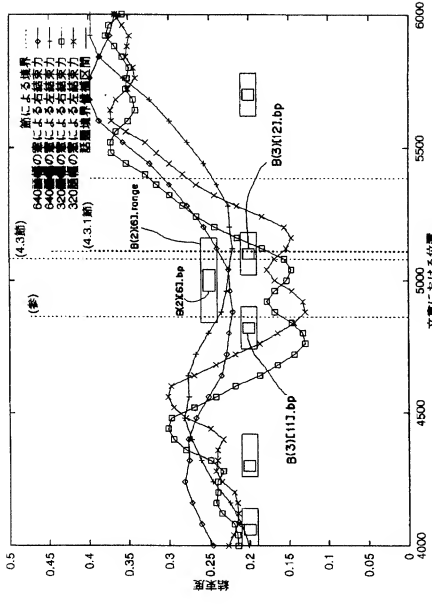
話題区間中に含まれている見出しを示す図

話題区間中に含まれている見出し

４．１　調査の概要
４．２　ネットワークアクセスのインタフェース
４．２．１　提言：１０年後のネットワークアクセスインタフェースはこうなる
（１）ネットワーク情報への多様なアクセス
（２）個人向けインタフェースを支えるエージェント技術
（３）セキュリティ・個人認証の今後
（４）機械翻訳と多国語

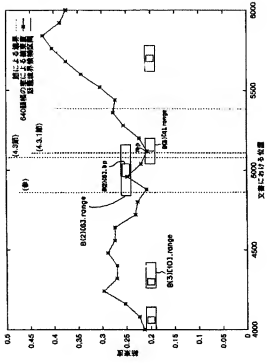
【図２７】

統合対象データを示す図



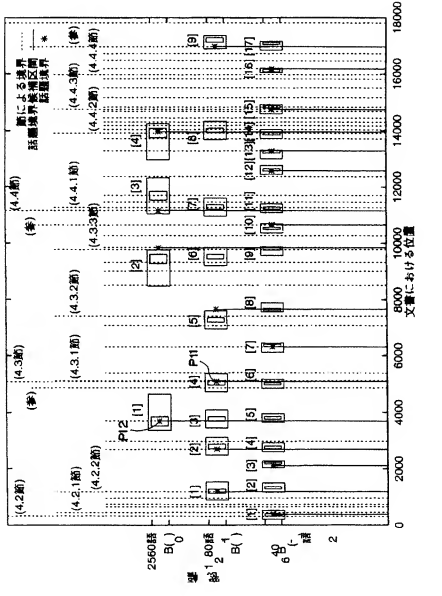
【図２８】

擬似データの作成方法を示す図



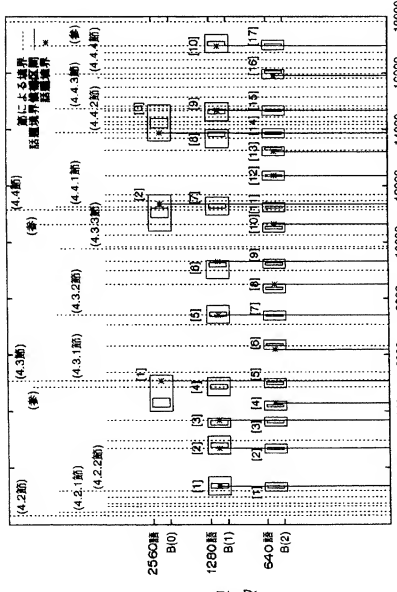
【図29】

話題構成の第1の認定結果を示す図



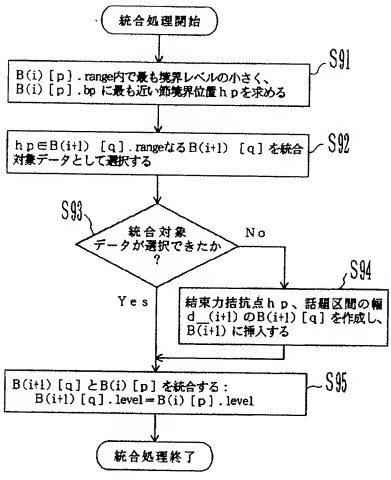
【図３０】

話題構成の第２の認定結果を示す図



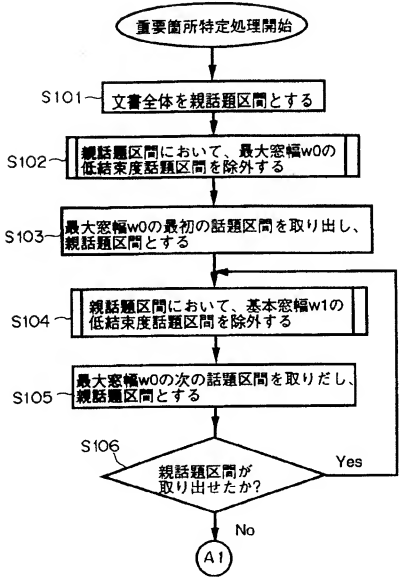
【図３８】

第２の統合処理を示す図



【図９】

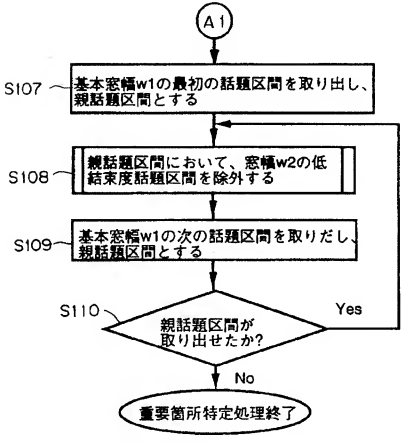
重要箇所特定処理のフローチャート
（その１）



【図４０】

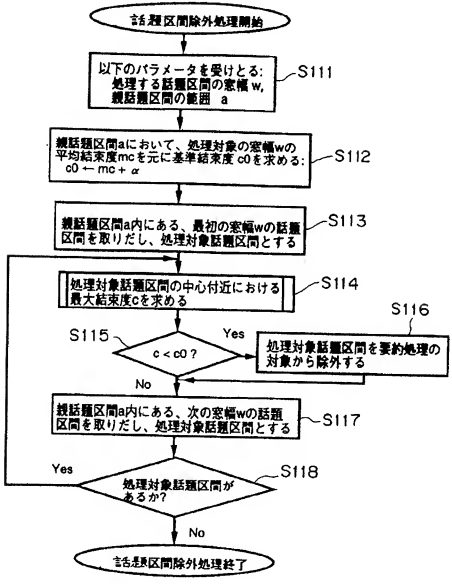
重要箇所特定処理のフローチャート

（その２）



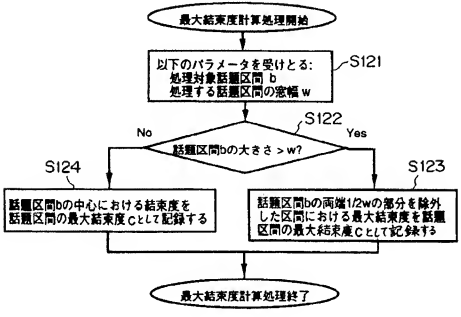
【図４１】

話題区間除外処理のフローチャート



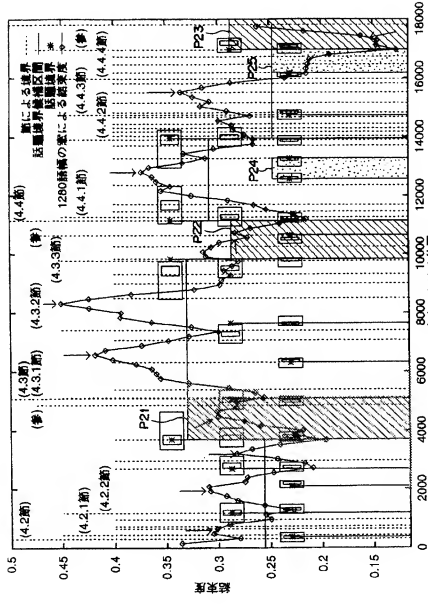
【図４２】

最大結束度計算処理のフローチャート



【図４３】

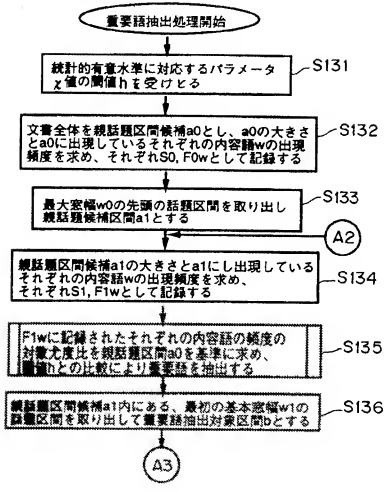
重要箇所の第１の特定結果を示す図



【図４４】

重要語抽出処理のフローチャート

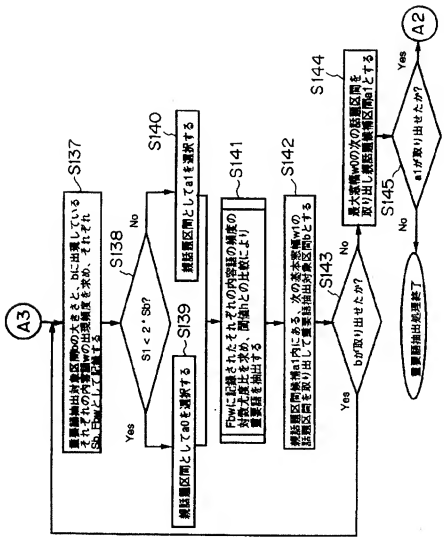
（その１）



【図４５】

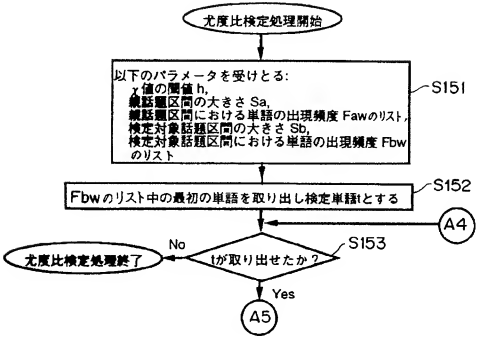
重要語抽出処理のフローチャート

（その２）



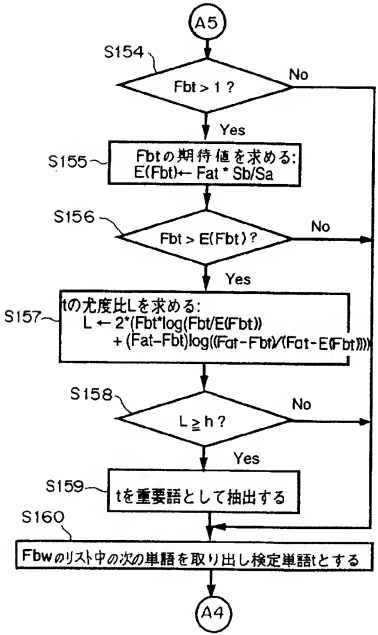
【図４６】

尤度比検定処理のフローチャート（その１）



【図４７】

尤度比検定処理のフローチャート（その2）



【図４９】

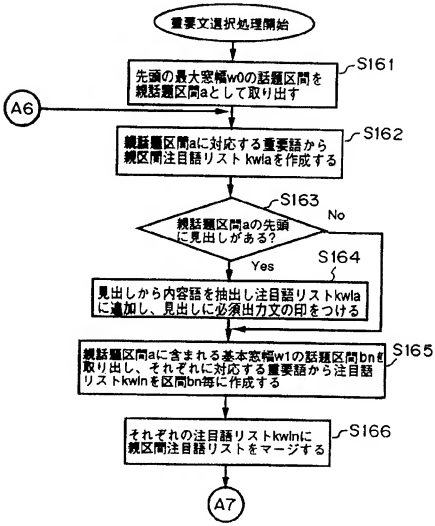
話題区間から抽出された重要語を示す図

話題区間から抽出された重要語

重要語 ↑	尤度比 L	区間内頻度 Fbt	親区間頻度 Fat
処理 / する	35.6	21	23
技術 /	34.7	38	53
な / る	28.4	34	49
調査 / する	27.3	15	16
アクセス / する	21.2	32	50
インターネット /	21.1	23	32
関わ / る	20.4	9	9
翻訳 / する	18.1	8	8
予想 / する	15.9	7	7
進展 / する	15.9	7	7
自然言語 /	14.7	9	10
インタフェース /	13.7	14	19
ネットワーク /	12.5	21	34
情報 /	11.6	51	106
量 /	11.3	5	5
家庭 /	11.3	5	5
パスワード /	11.3	5	5
重要 /	10.6	7	8
サービス / する	9.75	8	10
管理 / する	9.71	8	10
国語 /	9.08	4	4
母国語 /	9.08	4	4
資源 /	9.08	4	4
発展 / する	9.08	4	4
現状 /	8.66	6	7

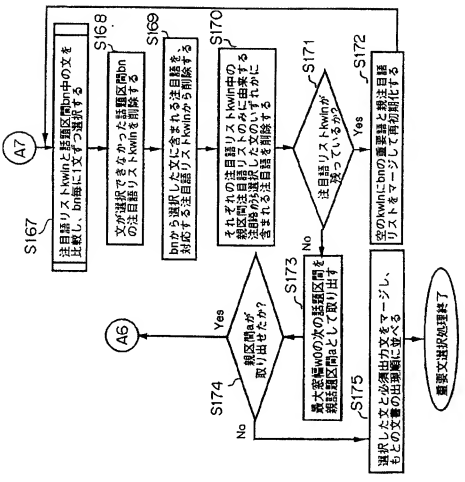
【図５０】

重要文選択処理のフローチャート（その１）



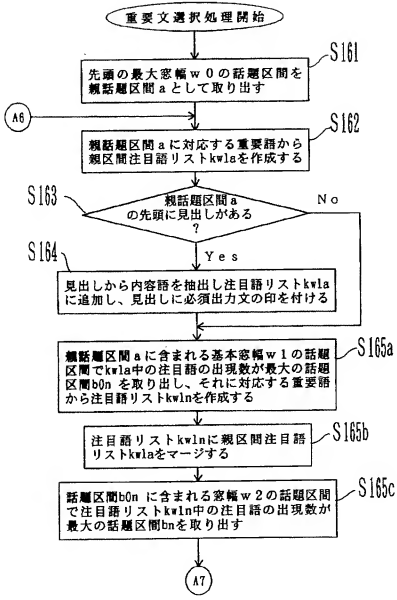
【図51】

重要文選択処理のフローチャート（その2）



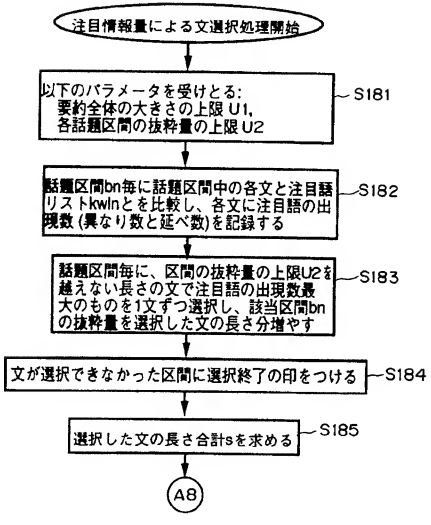
【図５２】

重要文選択処理の他のフローチャート



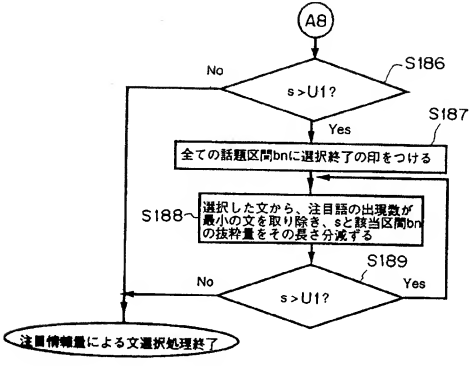
【図５３】

選択処理のフローチャート（その１）



【図５４】

選択処理のフローチャート　（その２）



【図55】

第1の発明の構成図

【図56】 第1の要約結果を示す図(その2)

4.3ネットワーク上の検索サービス
…
…また検索精度を高めるために、高頻度語は検索の対象としない、タイトルや見出しに含まれる語に重みをつける、などの工夫がなされている。
…
また、検索サービスが収集したページ数が膨大になるにつれて、ヒット数も膨大になってきたため、すばやく必要な情報を探すために、よりわかりやすい自動抄録作成技術が必要となる。…
…
tf・idf方式とは、単語に分割された文章の各単語の重要度を、その単語が文書中に出現する頻度tfと、その単語を含む文書が文書集合中に出現する頻度の逆数idfの積によってその単語の重要さを数値化する手法である。
…
[河合, 92]の研究キーワードのカイ二乗値から各キーワードの分類に対する得点を計算する場合に、シソーラス辞書から得られる抽象的な意味を得点に加える手法である。…
…

【図59】 第2の入力文書を示す図

SGML Type Document Managing Apparatus and Managing Method
Background of the Invention Field of the Invention

The present invention relates to an SGML (Standard Generalized Markup Language) document managing apparatus for allowing users to collaboratively create, edit, and revise a large SGML document sequence, such as a manual.

【図57】

第1の要約結果を示す図（その3）

4.4. 検索エンジン
…
[衛藤, 92]では、出現位置を記録する転分文字列索引において、字種に基づく黒台単位(漢字は文字、連続するかな2文字、かな文字と非かな文字の変化点などの設定、文字位置情報の昇順配列、低解像度黒台単位からの文字位置照合などを組み合わせることにより、専用ハードを用いなくてもソフトウエアで日本語の全文検索を高速に行えることを示した。
…
4.4.2. 有限オートマトンによる自然言語処理技術の動向
…
…しかし、パタンが有限オートマトンとして与えられた場合は、上記のトライのように決定と探索（そして、その接尾部分）が一意に定まらない為、失敗遷移が構成できない。…
…
…[Brill, 92]のタガーは、規則ベースでありながら、規則の自動獲得と適用 順序の学習により統計ベースのものと同等 レベルの耐性と精度が得られ、かつ、よりコンパクトであるという特長があるものの、規則が多くなるにつれ処理 速度が低下する。…
…

【図60】

第2の単語認定結果を示す図

SGML[SGML] Type[type] Document[document] Managing [managing] Apparatus[apparatus]
and Managing[managing] Method[method]
Background[background] of the Invention[invention] Field[field] of the Invention[invention]
The present[present] invention[invention] relates[relates]to an SGML[SGML]
(Standard[standard] Generalized[generalized] Markup[markup] Language[language]) document[document] managing[managing] apparatus[apparatus]for allowing[allowing] users[users]
to collaboratively[collaboratively] create [create], edit[edit], and revise[revise] a large[large]
SGML[SGML] document[document]sequence[sequence], such as a manual[manual].

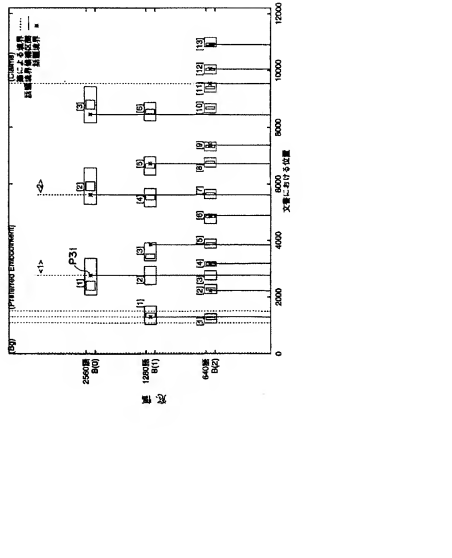
【図６１】

ストップワードを示す図

a, after, against, all, along, already, also, although, always, among, an, and, and/or, another, any, anywhere, are, as, at, be, because, been, before, before/after, being, belonging, belongs, below, between, both, but, by, can, cannot, corresponding, do, does, each, either, else, especially, even, every, for, forth, from, further, has, have, he/she, his/her, however, if, in, into, is, it, its, just, later, least, mainly, may, more, moreover, most, much, namely, next, no, not, of, on, once, one, only, or, other, others, otherwise, out, part, previous, same, should, since, so, some, someone, such, than, that, the, their, them, then, there, thereafter, thereof, these, they, this, those, through, thus, to, too, two, types, under, unless, unlike, until, up, usually, was, well, were, what, when, where, whereby, wherein, whether, which, while, who, whole, whose, why, will, with, with/without, without, yes

【図６３】

話題構成の第３の認定結果を示す図



【図６４】

重要箇所の第2の判定結果を示す図

【図６７】

第２の要約結果を示す図（その３）

What is claimed is:

...

15. The document **managing** apparatus as set forth in **claim** 14, further comprising:**;**br**;**edited result output **means** for outputting the electric document to be edited corresponding to the edited result to **said database means** when the determined result of **said** editing consistency **examining means** is satisfied.

...

【図６５】

第２の要約結果を示す図（その１）

SGML Type Document Managing Apparatus and Managing Method

...

... In addition, SGML documents have been widely used for fields that handle large and long-life documents such as network communications, electronic trading, and databases such as electronic libraries. ...

...

The present invention is mainly intended to provide an SGML type document managing apparatus and an SGML type document managing method that allow collaborative creating and editing works to be effectively performed. ...

...

Brief Description of Drawings

...

Fig. 2A is a block diagram showing the structure of the client 2 of the information processing system in the case the structure shown in Fig. 1 is accomplished by software.

...

【図66】

第2の要約結果を示す図（その2）

The CPUs 100-1 and 100-2 of the **client** 2 and the **server** 3 execute **software** that is stored in the main **storing units** 400-1 and 400-2, the **software** being read from the **auxiliary storing unit** 200-1 or the **input/output units** 500-1 and 500-2, or the **software** being obtained from the connecting **network** through the **network connecting units** 300-1 and 300-2, **respectively**....

...

At **step** S71, the SGML document **accessing** unit 30 **adds** the fist item of the **repeatable model group** without a **head**, namely the **removable** instance **number** and a generic identifier/model **group** with a occurrence indicator, and "(" to the **top** of the **extracted** extended **content model**.

...

At **step** S160, the SGML document editing unit 10 **removes** the **identifiers** (the document **identifier** and the **element identifier**) of the **element** stored in the **register** from the deleted **element attribute** (deletions) of the parent **element** of the **paste destination**.

...